GUIDELINES FOR STATISTICAL ANALYSIS OF OCCUPATIONAL EXPOSURE DATA

FINAL

by

IT Environmental Programs, Inc.
11499 Chester Road
Cincinnati, Ohio 45246-0100

and

ICF Kaiser Incorporated
9300 Lee Highway
Fairfax, Virginia 22031-1207

Contract No. 68-D2-0064
Work Assignment No. 006

for

OFFICE OF POLLUTION PREVENTION AND TOXICS
U.S. ENVIRONMENTAL PROTECTION AGENCY
401 M STREET, S.W.
WASHINGTON, D.C. 20460

August 1994

DISCLAIMER

This report was developed as an in-house working document and the procedures and methods presented are subject to change. Any policy issues discussed in the document have not been subjected to agency review and do not necessarily reflect official agency policy. Mention of trade names or products does not constitute endorsement or recommendation for use.

CONTENTS

FIGURES

TABLES

ACKNOWLEDGMENT

Many individuals and organizations have been helpful in developing this report; for these contributions the project management extends its sincere gratitude.

Mr. Paul Quillen and Ms. Breeda Reilly were the EPA Project Officers and Ms. Cathy Fehrenbacher was the EPA Work Assignment Manager. Mr. Thomas Corwin, IT Environmental Programs, Inc., was the Project Director and Mr. Edwin Pfetzing the Project Manager. Mr. Robert Goodman, IT Environmental Programs, Inc., assisted in the preparation of the report. Ms. Nora Zirps was the ICF Project Manager. Dr. Erwin Hearne and Mr. Bruce Allen, K.S. Crump Division of ICF Kaiser developed the statistical methodology for the report. Extensive review of and comment on the guidelines was made by Drs. Rick Hornung, Larry Elliott, Steve Ahrenholz, David Utterback, and Thurman Wenzel, NIOSH; and Elizabeth Margosches and Gary Grindstaff, EPA.

Peer review was provided by Dr. Neil C. Hawkins, Dow Chemical Company; Mr. Keith A. Motley, OSHA; Dr. Stephen M. Rappaport, University of North Carolina; Col. James C. Rock, U.S. Air Force; and Dr. Steve Selvin, University of California, Berkeley.

INTRODUCTION

The purpose of these guidelines is to establish a consistent approach to handling the wide variety of occupational exposure data available for preparing occupational exposure assessments in support of risk assessments. It provides guidance in the characterization of broad ranges of job groups with similar exposures, calculation of descriptive statistics (where appropriate) and treatment of uncertainties, assumptions, and biases in the data. It is designed to be used by engineers in the Office of Pollution Prevention and Toxics (OPPT), with some assistance from industrial hygienists and statisticians. The procedures described provide a systematic methodology for performing an occupational exposure assessment based upon the types of data which are most commonly available for such analyses. Methods used by OPPT's Chemical Engineering Branch (CEB) to prepare assessments of occupational exposure and environmental release are presented in the CEB Engineering Manual (IT, 91). These guidelines are a supplement to the CEB Engineering Manual intended for use with recently collected data. It should be noted that these guidelines are not intended to provide recommendations for performing additional monitoring of exposure or for determining compliance with regulatory standards. If this is the goal, the reader should consult other references such as Hawkins (91) and Patty (81), etc.

A. Types of Occupational Exposure Monitoring Data

Monitoring data usually consist of area samples, personal inhalation samples or dermal samples. Area samples are collected to represent the airborne concentration of a chemical in a specific location at a facility. Personal samples are collected to represent a worker's inhalation exposure during a specified time period; for example, peak, ceiling, short-term, and full-shift samples. Peak or ceiling samples are typically collected instantaneously through continuous monitoring or for 15 minutes or less. Short-term samples are collected over a designated period, typically less than 2 hours. Full-shift samples are collected to represent a worker's inhalation exposure over an entire work shift and may be composed of a single sample or consecutive short-term samples. Dermal samples are collected to represent a worker's dermal exposure to a given chemical over a portion of the body which has been in contact with the chemical. Exposure data collected for each type of exposure should be separated and statistical analyses conducted separately.

Biological monitoring may also be used to determine an employee's overall exposure to a given chemical by measuring the appropriate determinant in biological specimens collected from exposed workers at the specified time. While biological monitoring provides information complementary to air monitoring, interpretation of data can be difficult due to variability in the physiological and health status of the individual, exposure sources, individual life style, analytical errors, etc. If biological monitoring data are

available, this fact should be noted in the exposure assessment. This report does not address biological monitoring but focuses on air monitoring data collected to assess inhalation exposure.

For the purposes of this report, three broad categories of occupational exposure data are considered:

- Type 1 data consist of measurements for which all important variables are known. The data consist of studies that contain individual measurements and include all backup and ancillary information (e.g., analytical method, limit of detection, sampling duration, type of sample taken, job tasks, etc.).

- Type 2 data consist of measurements where important variables are not known but for which assumptions can be made for their estimation. The data consist of individual monitoring measurements, but backup and ancillary information are inconsistent.

- Type 3 data consist of measurement summaries, anecdotal data, or other data for which the important variables are not known and cannot be estimated. Individual monitoring measurements are typically not available.

These categories were developed for use with these guidelines; judgment is used in determining the type(s) of data available. Examples and additional information on the categories are provided beginning with Step 10.

Once satisfied that the data have been properly collected for the objective of the study, the primary determinant of the confidence one can place in the analysis is the sample size. Every effort should therefore be made to collect and analyze every available piece of data. Because the size of the data set being analyzed has a large effect on the confidence that can be placed in the analysis, the methodology set forth in these guidelines allows the combination of similar data sets based on statistical tests. The traditional categorization of data by the industrial hygienist or engineer is supplemented by statistical analysis of the categorization; the goal is identification of groups of data that are as large as possible and describable by standard statistical distributions (lognormal and normal).

B. Types of Occupational Exposure Assessments

There are various types of exposure assessments performed by OPPTs' CEB. The main distinction between them is the level of effort expended in collecting data. Regardless of what type of data are obtained, however, the CEB engineer should review the level of detail required in the exposure assessment and try to provide the best and most complete analysis of the available data.

The following are examples of the program areas and types of exposure assessments performed by CEB:

- New Chemicals Program. An initial screening assessment is performed with a goal to determine the high end and central tendency exposures, generally using available information and information submitted in the Premanufacture Notification (PMN). In reality, these estimates are more likely to be bounding (e.g., overestimates of) exposure, due to lack of information. If there are concerns for worker exposure, the initial assessment is refined as the case progresses through the review process. However, due to lack of data on these new chemicals which have not yet been commercialized, this often involves the use of modeling or surrogate data, rather than analysis of actual data on exposure to the substance of concern.

- Chemical Testing. A preliminary exposure assessment is completed to determine the bounds of potential occupational exposure for chemical testing candidates. This exposure assessment is refined as the case progresses and additional information is gathered. Since these are "existing" chemicals, there may be some exposure data available on the specific substance. These chemicals may be referred to CEB through the Interagency Testing Committee (ITC).

- Existing Chemicals. An exposure assessment may be an initial screening which is used to help determine if further work is needed on the case. If so, a more detailed exposure assessment including the range of potential exposure, measure of central tendency, uncertainty, etc. is completed for the population(s) of concern. A risk assessment is performed; if risk management action will be taken the exposure assessment may be revised to include additional information or to cover additional uses, etc. For some cases monitoring studies will be conducted to determine workplace exposure levels. An evaluation of controls may also be needed.

C. Variability in Occupational Exposure Data

It is rare to find studies of occupational exposure based on a statistical approach to providing representative information for an individual facility; it is even less likely to find such a study that represents a particular industry subsector or group of facilities. While random sampling (i.e., monitoring exposure to a group of workers in a random fashion) is preferred, "worst-case sampling" (i.e., monitoring the individual with the highest exposure) during a 1- to 3-day sampling campaign is common industrial hygiene practice for compliance with regulatory standards. However, sampling programs are being used that promote exposure monitoring and periodic surveillance (Damiano, 89; Hawkins, 91).

Even in statistically-selected, well-done studies, there may be high variability in the characterization of worker exposure. Measurements at a plant made over a period of no more than a few days may be all that are available to characterize exposures over an entire year or a period of years. Seasonal variability, interday and intraday variability, and changes in the process or worker activities can cause the exposure to vary from that measured on a single day. Temperature changes can affect evaporation rates, and seasonal changes in natural ventilation affect exposure. Sampling methods and time periods can also vary. Seldom can all these variables be measured and accounted for. However, if

important variables are identified and quantified, it is hoped the influence of less important variables on the overall measure of central tendency will be minimized. Variables that may not be obvious may also affect variability among plants in the same industry category. Variables such as the age of the plant, the age of the control equipment, whether the plant is in a volatile organic compound (VOC) nonattainment area, and operation and maintenance (O&M) practices at the plant should be investigated.

When analyzing sample data, it is important to understand the sources of variation in exposure sample results that combine to create the observed variability (Patty, 81). The size of the variation may be a function of both the exposure levels and the measurement method. Both random and systematic errors should be considered.

Random variations in workplace exposure levels can result in intraday variations, interday variations, or variations in exposures of different workers within a job group or occupational category (Patty, 81). Variability in the measurement procedure can be caused by random changes in pump flow rate, collection efficiency, or desorption efficiency. It is important to realize that random variation in real workplace exposure levels will usually exceed measurement procedure variation by a substantial amount, often by factors of 10 or 20 (Patty, 81; Nicas, 91).

Systematic variations in the determinant variables affecting workplace exposure levels will lead to systematic shifts in the exposure results. Variability in worker exposure levels reflects changes in worker job operations during a work shift or over several days, production process changes, or control system changes. Systematic errors in the measurement procedure can result from mistakes in pump calibration, use of sampling devices at temperatures or altitudes substantially different from calibration conditions, physical or chemical interferences, sample degradation during storage, internal laboratory errors, and interlaboratory errors (Patty, 81). These errors may be identified and their effects minimized with the use of quality assurance programs (EPA, 92). Specific variables (parameters) that can affect occupational exposure measurements are more fully discussed in Step 4.

It is also important to ascertain the objectives of the monitoring study to identify potential biases in the data. For example, if the objective was to sample only well-controlled facilities, then the results would probably not represent the exposure in the industry as a whole. If the monitoring resulted from worker complaints, then exposures may not represent typical exposures. If the monitoring was conducted to evaluate engineering controls or as a preliminary screening of exposure, the results may not represent actual employee exposure. It is important that all potential variables be identified and evaluated.

D. Organization of This Report

Following the introduction is a 19-step procedure for statistical analysis of occupational exposure data. Figures 1 to 3 present flow diagrams outlining these procedures. Each numbered step in these figures is explained separately. Steps 1 through 6 are presented in Figure 1 and give the actions necessary to prepare a preliminary exposure matrix. Steps 7 through 14 are presented in Figure 2 and give the actions necessary to prepare a completed exposure matrix from the preliminary exposure matrix including preparation of a non-statistical report on Type 3 data. Steps 15 through 19 are presented in Figure 3 and

relate to the statistical analysis of Type 1 and 2 data and the presentation of the results. An example is used throughout the 19 steps to better explain the techniques used in the guidelines. The data used in the example are based on real data, but have been altered where necessary to emphasize particular points in the guidelines.

These guidelines present rather sophisticated approaches for statistical analysis of occupational exposure data. Nonstatisticians may require training or the assistance of a statistician in order to properly understand and use the guidelines. The development of software as a companion to the guidelines could be useful in guiding the user through the analyses and in incorporating more complex calculations for certain nondefault procedures discussed in Appendix B.

A bibliography of references pertinent to occupational exposure analysis is also provided. Appendix A presents a spreadsheet matrix for the example data set. Appendix B presents background information on the methodology available to statistically analyze the data. Appendix C presents a listing of currently available computer software for the statistical analyses.

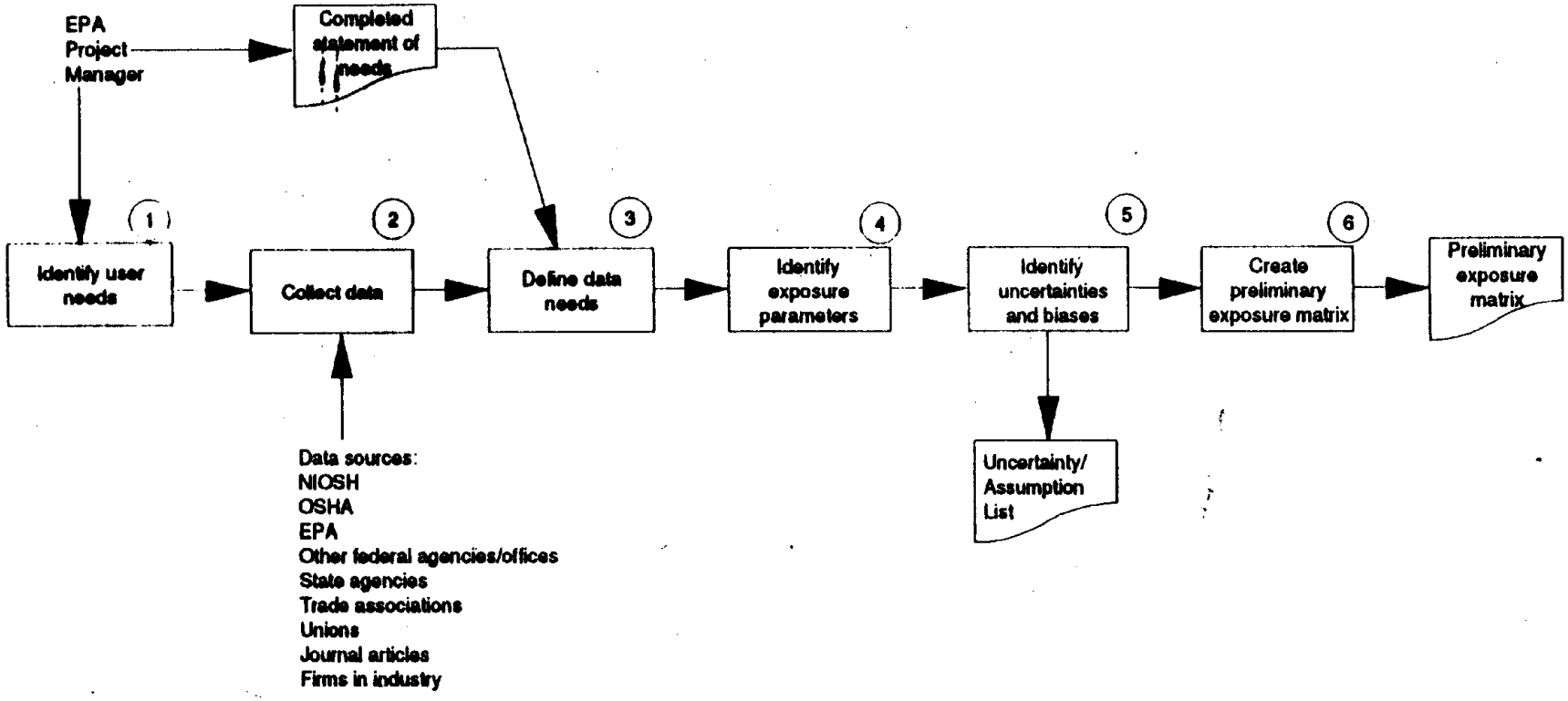

Figure 1. Flow Diagram for Creation of Preliminary Exposure Matrix.

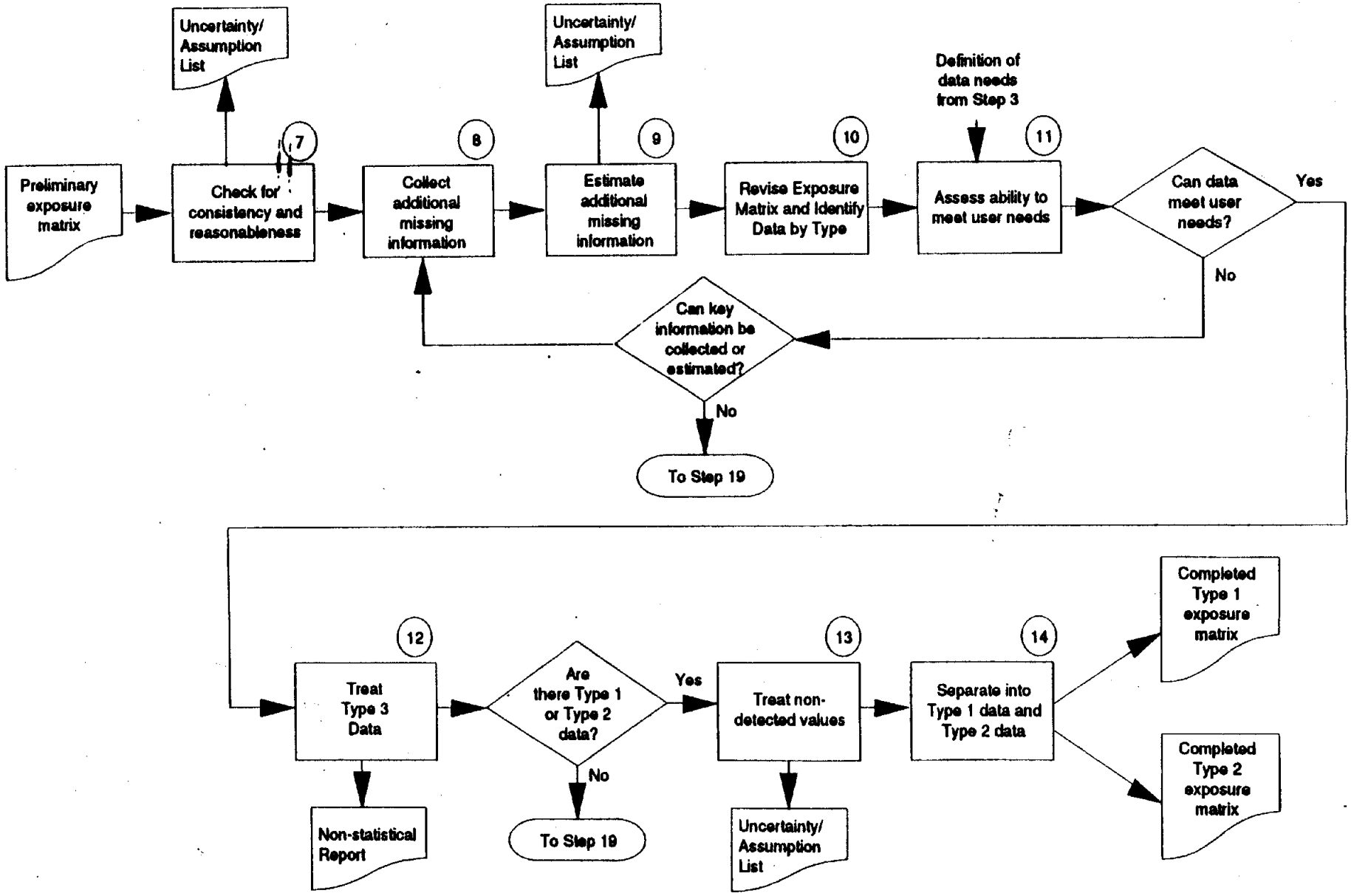

Figure 2. Flow Diagram for Creation of a Completed Exposure Matrix.

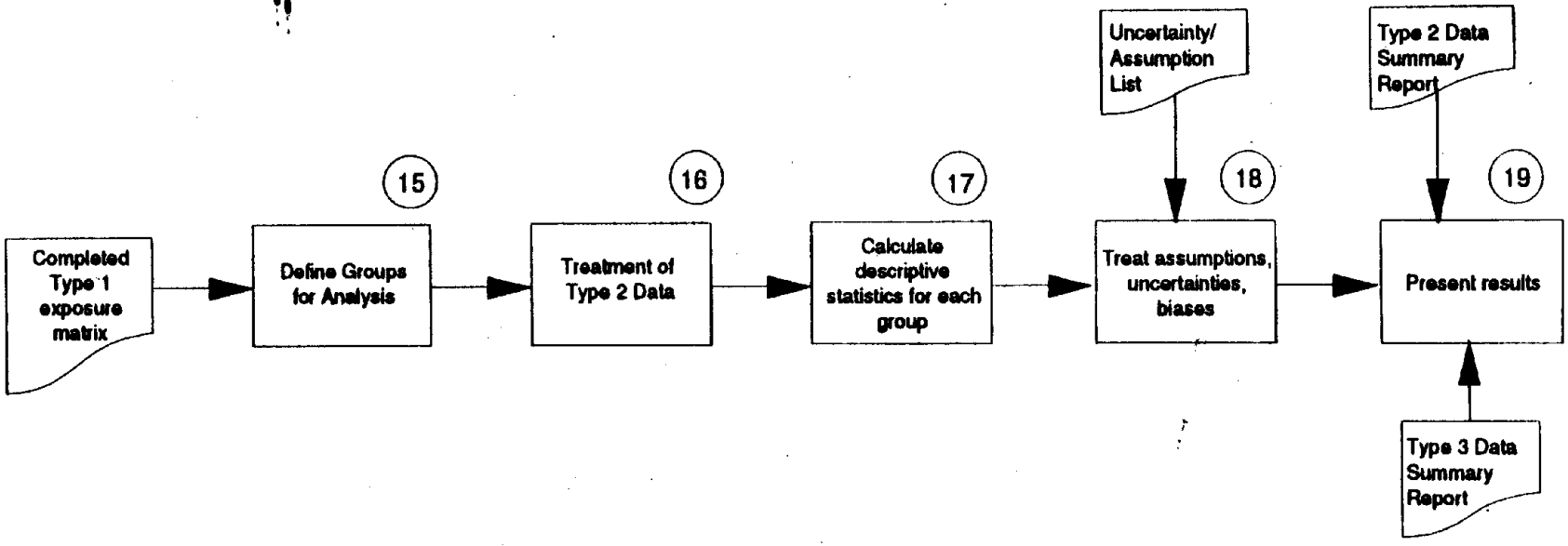

Figure 3. Flow Diagram for the Statistical Analysis of Type 1 and Type 2 Data.

STEP 1: IDENTIFY USER NEEDS

The first step in an exposure assessment is to identify the needs of those using the information, usually in some form of risk assessment activity. The user is typically the project manager for the chemical under review. This step initially identifies the data requirements of the assessment so that resources can be used most effectively to collect pertinent data.

The level of detail required in an exposure assessment depends on the scope of the risk assessment activity it supports (EPA, 87). If the purpose of the analysis is merely to screen a new chemical for potential problems, a much less rigorous bounding estimate of exposure will often be prepared. These analyses are useful in developing statements that exposures are "not greater than" the estimated value. However, to support a detailed risk assessment, an in-depth presentation of potential exposures must be prepared. It is also necessary to know if the end user is interested in a particular demographic group, route of exposure, frequency and duration of exposure, industry, exposure period, or other variable. For example, if the chemical is of concern because of possible reproductive effects for women of childbearing age, then every effort should be made to gather information on the exposure of this demographic group. Information needs also depend on the specific health hazards identified for the chemical. Some of the information needs that may be identified include:

- Mean, standard deviation
- Geometric mean, geometric standard deviation
- Range of exposures, confidence intervals
- Duration of exposure (hr/day and days/yr)
- 8-hour time-weighted average (TWA)
- Peak exposures
- Time period (i.e., particular year, 1989)
- Cumulative exposure over time, lifetime average daily exposure (for possible use in risk assessment)
- Probability of excursions or exposure during upsets or emergency release
- Uncertainties associated with the data and assumptions used in analyzing the data

The objectives of the exposure assessment must be defined using information obtained from the "user," typically the project manager for the chemical under review. To assist in this process project managers should be contacted initially to discuss the data requirements of the assessment and asked to complete a "statement of customer needs" form for exposure assessments which are not typical new chemical-type assessments. When this form (shown in Figure 4) is returned, it will be of value in Step 3 to more completely define user needs.

Since health effects data are often gathered to prepare the hazard assessment in parallel to the occupational exposure assessment, good lines of communication with the project manager and those preparing the hazard assessment will facilitate information exchange regarding potentially changing assessment needs. For example, as new health effects are defined, the exposure data classification or level of detection required of the analytical methods used may need to be changed. For example, if chronic

health effects are identified, generally long-term exposures are of interest, while peak or short-term exposures are of interest for acute health effects. Timely communication will minimize the changes that need to be made as well as the need for further data collection.

EXAMPLE

The example shown below will be used throughout this report to illustrate how the statistical analysis proceeds.

The example chemical is a colorless gas whose primary use is polymerization to make various elastomers. Recent chronic oncology studies indicate that the chemical is carcinogenic in mice. The present OSHA Permissible Exposure Limit (PEL) is 1,000 ppm as an 8-hour TWA, but the American Conference of Governmental Industrial Hygienists (ACGIH) recommended a revised Threshold Limit Value (TLV) of 10 ppm as an 8-hour TWA.

The project manager identified two general needs for the exposure assessment. First, the exposure assessment was needed to do a preliminary risk assessment for all worker exposures to the chemical. Second, it was needed as a baseline to estimate the technological feasibility and cost of reducing worker exposure to target levels of 10 ppm, 1 ppm, and 0.1 ppm. An example statement of needs form for the example chemical is shown in Figure 4.

Figure 4. Statement of Needs

Statement of Customer Needs for CEB Engineering Assessments

Requester: *Sally Jones, Project Manager* ***Date of Request:*** *2/20/94*

The purpose of this form is to gather information on customer needs to be used in developing a CEB engineering assessment. Please note that all identified needs may not be met due to data limitations, resource constraints, etc. What with multiple customers of CEB assessments, it is suggested that the form be completed by the individual who will be using the specific type of information provided by CEB.

Return completed form to: *John Smith, CEB Engineer* ***Phone:*** *260-1234*

Section 1. General Information

A. Please indicate the origin of the case and chemical/use cluster, etc. (e.g. RM 2 analysis for hydrazine): RM2 analysis for example chemical.

B. What are the purpose and goals of the CEB assessment and the project? Develop assessment of occupational exposure to the example chemical.

C. What are the approximate completion dates for the CEB assessment and for the project? CEB assessment is due April 4, 1994

D. Please identify the health effects of concern (e.g. carcinogenicity, neurotoxicity, liver effects, reproductive effects, sensitization, etc.): Carcinogenicity

E. Please identify the environmental effects of concern : NA

F. Please identify any specific data, sources, references, or personal contacts you would like CEB to research: NIOSH and OSHA data.

G. When do you need to have an estimate of CEB extramural resources (if any) for this project? NA

Section 2. Occupational Exposure Assessment ☐ *Not Needed*

A. CEB will estimate number of workers exposed for each industry segment of interest. Identify any special population characteristics of interest (e.g. gender, etc.): Total number of workers potentially exposed, and population potentially exposed during monomer and polymer production.

B. Identify specific industry segment(s) of interest (e.g. manufacture, processing and end uses; only spray coating application end uses, etc.): Monomer and polymer production.

C. Indicate ✓ which types of exposure are of interest:

✓ Inhalation exposure ☐ Dermal exposure

☐ Other (e.g. ingestion): _____

D. Identify which worker activities are of interest (e.g. the assessment need only address textile dye weighers): All worker activities associated with monomer and polymer production.

E. Indicate ✓ the preferred characterization for duration and frequency of exposure:

☐ Short-term exposure (e.g. peak exposure, maximum 15-minute exposure, etc.), for acute health effects. Identify specific requirements: _____

✓ Long-term exposure (e.g. annual average exposure, lifetime average daily dose, etc.), for chronic health effects. Identify specific requirements: annual average exposure and lifetime average daily dose.

✓ Frequency of exposure (days/yr)

✓ Cumulative exposure over time (e.g. days, months, years): days, months, and years are of interest

☐ Other: _____

G. CEB will attempt to provide a measure of central tendency, and a high end Potential Dose Rate (PDR), identify assumptions made, and characterize uncertainty, as data and methodologies allow. Identify any specific needs (e.g. specific statistical descriptors, etc): Statistical descriptors of geometric mean, arithmetic mean, geometric standard deviation, arithmetic standard deviation, the distribution of the data, and a graphic presentation of the data are preferred.

H. Please identify any other special needs for the occupational exposure assessment: Estimate of the technical feasibility of controlling exposure to 10 ppm, 1 ppm and 0.1 ppm.

Section 3. Process Information ✓ *Not Needed*

A. Are there specific industrial segments (e.g. manufacture, processing into a coating, end use as a paint in automotive application) you would like process information for?

1. _____
2. _____
3. _____
4. _____

B. Please specify the information you would like CEB to provide:

☐	Number of sites	☐	Days/yr
☐	Throughput (kg/site-day)	☐	Process Description
☐	Flow Diagram	☐	Other (please specify) _____

_____

_____

Section 4. Environmental Release Assessment ✓ *Not Needed*

A. CEB will provide estimates of environmental release (i.e. kg/site-day or kg/yr) for manufacture, processing and end use operations. Indicate any specific industry segments of interest or special data needs: _____

_____

_____

_____

B. Indicate ✓ which types of releases are of interest, and indicate any special needs:

☐	Water releases	☐	Air releases
☐	Landfill releases	☐	Incineration releases
☐	Other: _____		

Special Needs: _____

_____

C. CEB will attempt to provide descriptors for release assessments, identify assumptions made, and characterize uncertainty, as data and methodologies allow. Identify any specific needs: _____

_____

_____

_____

Section 5. Pollution Prevention Assessment (PPA)/Occupational Exposure Reduction Assessment (OERA) *✓ Not Needed*

Are there specific industrial segments you would like CEB to provide an assessment of pollution prevention opportunities and/or occupational exposure reduction for?

□ PPA **□ OERA** **□ Both**

1. _____
2. _____
3. _____

Section 6. Other Information Needs *✓ Not Needed*

Please identify other information, analysis or data needed, and the rationale for requiring the information: _____

_____

_____

Customer Contact (e.g. Project Manager):

Sally Jones, Project Manager CCD 260-2345 2/20/94

(Name) (Division/Branch) (Telephone) (Date)

STEP 2: COLLECT DATA

Once the data requirements of the assessment are preliminarily identified, the next step is to collect the monitoring data that will be used in the analysis. It is important to obtain information on all variables relating to the measured values, such as the collection method, number of workers exposed, duration of the sampling, etc. Step 4 contains a listing of parameters that may affect exposure. The more data that are identified and collected, the better the analysis will be. Therefore, it is important to ascertain at the beginning of the project that all possible sources of data have been checked.

Typical sources of exposure monitoring data include the National Institute for Occupational Safety and Health (NIOSH), the Occupational Safety and Health Administration (OSHA), the Environmental Protection Agency (EPA), other federal agencies or departments, state agencies, trade associations, unions, journal articles, and individual companies in the industry.

A. Obtaining Data From NIOSH

For existing chemicals that have been studied by NIOSH, Health Hazard Evaluations (HHEs) and Industry Wide Surveys (IWSs) usually represent the largest body of complete and extremely well documented data. NIOSH reports usually include most of the information necessary to fully classify data. In cases where the chemical of interest was not the primary reason for the NIOSH report, but rather only measured as a secondary chemical, information may have to be filled in by direct contact with the inspector. In addition, it may also be necessary to confirm the presence of the chemical in all areas monitored if a large quantity of nondetected values are recorded. Since HHEs are generally done in response to a complaint regarding a specific chemical, the data may not be random in selection. IWSs tend to be well selected to represent an industry, but may be biased if only well controlled facilities were monitored. NIOSH Control Technology Assessment reports are developed to identify and evaluate appropriate control measures and may be biased toward facilities that are well-controlled. Contact with NIOSH can usually identify any potential biases. NIOSH tends to take many samples per visit as contrasted with OSHA which typically only takes a few measurements.

In general, NIOSH inspectors are easy to locate and will have worked on more than one of the surveys, so that multiple information can be gathered from each contact. Where contact cannot be made, it is usually acceptable to assume that the NIOSH collection and analytical method recommended at the time was used to collect the data. NIOSH may also have unpublished data or studies that are in progress; contact with NIOSH personnel who have been or are working on the chemical can thus result in additional unpublished monitoring data. The best source of NIOSH reports is the NIOSHTIC data base, which is available through DIALOG or on computer disk. In addition, the NIOSH Publications Catalog can be manually reviewed to identify useful reports. It may also be useful to obtain up-to-date published and unpublished information available on microfiche and hardcopy from NIOSH. Data may be obtained from:

U.S. Department of Health and Human Services
National Institute for Occupational Safety and Health
Robert A. Taft Laboratories
4676 Columbia Parkway
Cincinnati, Ohio 45226
(800) 35-NIOSH

B. Obtaining Data From OSHA

The largest number of measurements for an existing chemical is generally located through accessing the OSHA National Health Sampling Results by Inspection (OSHA report: OHR 2.6). These data can be obtained by written request to:

U.S. Department of Labor
Occupational Safety and Health Administration
Director, Office of Management Data Systems
Room N3661
200 Constitution Ave., N.W.
Washington, D.C. 20210
(202) 219-7008

Information provided for each facility includes company name and address, SIC code, inspector code, OSHA office, date and reason for visit, job title, exposure value, number of similarly exposed workers at the time of the inspection, and type of exposure (peak/8-hour TWA, personal/area). No information is provided on controls, type of process, monitoring method, concentration of chemical in process, or demographics of the exposed workers. The sampling and analytical method and limit of detection may not be available. Where the sampling and analytical method cannot be ascertained, it is usually acceptable to assume that the method used is that specified by OSHA in the OSHA Technical Manual at the time the survey was used (OSHA, 90). The methods specified in this publication are in most cases from either the NIOSH Manual of Analytical Methods (NIOSH, 84) or the OSHA Manual of Analytical Methods (OSHA, unpublished). Unlike NIOSH, OSHA usually collects only one or two samples per chemical during each inspection. In many cases, the job title or SIC may uniquely define the use of the chemical (e.g., degreaser operator or SIC 7216, Dry Cleaning Plants), but most data require that some assumptions be made for categorization. In addition, the data may include large quantities of nondetects and SIC codes may be inconsistently applied. If time and budget permit, it is best to contact the OSHA inspector. Because the inspector at the local OSHA office must be called and few summaries are from the same inspector, this process can be time consuming. Also, inspectors may be difficult to locate, files may be stored away, or the inspector may not remember details of the facility. Many states (23 to date) operate their own OSHA State Programs which must be "at least as effective as" the federal program. However, these State plans have historically not had data in this OSHA data base. OSHA's Publication Catalog can also be reviewed, and up-to-date information (including NIOSH studies) may also be available from:

OSHA Technical Data Center
Department of Labor
200 Constitution Avenue, N.W.
Room H-2625
Washington, D.C. 20210
(202) 219-7500

C. Other Sources of Data

Monitoring data may also be available from previous and ongoing EPA studies. Previous reports done by OPPT (formerly OTS) may contain occupational exposure data. Usually the data will have been summarized and the primary data will have to be obtained separately. It is important to obtain primary data to avoid the duplication of data from other sources. Information submitted under Sections 4, 8(a), and 8(d) of TSCA may be useful in preparing the exposure assessment. Non-confidential information submitted under TSCA may be obtained through the TSCA Non-Confidential Information Center at (202) 260-7099. The Office of Air Quality Planning and Standards (OAQPS) may have collected some exposure data through the use of Section 114 letters. Information about OAQPS Section 114 letters can be obtained by contacting the Emissions Standards Division at (919) 541-5571.

Other federal agencies or departments may have collected exposure data. For example, the Army and Air Force have monitoring data on workers in a wide variety of job categories. These data may be obtained by contacting the following departments:

Army:
Assistant Secretary of the Army
(Installations, Logistics and Environment)
Attn: SAILE (ESOH)
1 110 Army Pentagon
Washington, D.C. 20310-0110
(703) 614-8464

Air Force:
HQ AFMOA SGPA (BEES)
170 Luke Avenue
Bolling AFB
Washington, D.C. 20332-5113
(202) 767-1731

MSHA:
Mine Safety and Health Administration
Metal/Nonmetal, Division of Health
4015 Wilson Blvd.
Arlington, VA 22203-1984
(703) 235-8307

Mine Safety and Health Administration
Coal, Division of Health
4015 Wilson Blvd.
Arlington, VA 22203-1984
(703) 235-1358

State environmental and occupational safety agencies concerned with both environmental protection and worker health may have monitoring data. This is especially true if there is a concentration of the industry under study in a state.

Trade associations often collect and evaluate monitoring data from their members. In many cases the association may not allow access to the primary data and will provide only summaries of the data, thus limiting its usefulness. Even if the data cannot be incorporated in the direct analysis, however, it can be used for comparison with the results of other analyses. An extensive listing of trade associations is contained in the Encyclopedia of Associations (Koek, 88).

Unions often are the driving force behind the investigation of a particular chemical. In such cases they may have obtained exposure measurements from companies with which they have contracts. Direct contact with the union in question is the best method to obtain these data.

Data may also be identified from journal articles. On-line data bases that can be useful to identify exposure data include BIOSIS, CA Search, EMBASE, Enviroline, Medline, NIOSHTIC, NTIS, and Pollution Abstracts. These sources almost never present the primary data and the necessary ancillary information, so the author will usually have to be contacted if primary data are necessary.

Finally, if plant visits are being conducted or plants are being contacted to provide information for the study, they may also be asked to voluntarily provide monitoring data. Such contacts are of course limited by Office of Management and Budget (OMB) oversight under the provisions of the Paperwork Reduction Act. Plants may also be surveyed in the form of OMB approved questionnaires or telephone surveys.

EXAMPLE

For the example chemical, worker exposure data were obtained from NIOSH, OSHA, a previous contractor report for EPA, and the union representing workers at several facilities. The data were generally not primary monitoring results but only summaries of the data giving means and number of samples for ranges (i.e., Type 3 data). The user needs identified in Step 1, however, called for the types of results only available by analysis of Type 1 data. Therefore, new monitoring data had to be collected for the industry. The available and new data form the basis for the analyses shown in the example in the following steps.

STEP 3: DEFINE DATA NEEDS

By the time the initial data collection has been finished, the completed "statement of needs for occupational exposure assessment" form (Figure 4) should have been received from the project manager. This form and any other information provided should be used to formally define the data needs of the assessment. A preliminary determination should be made by the CEB engineer as to whether the existing data are "in the ball park" or if significant changes in data collection resources or expectations of the project manager are needed. A more detailed assessment of whether the user needs can be met will be made in Step 11.

If it is apparent that the exposure data are inadequate to meet the needs set forth in the statement of needsy form, then the CEB engineer should inform the project manager that expectations should be modified to match the existing data or outline approaches and resource implications to meet those needs.

It is important to be responsive to requests for specific statistics in the assessment. For instance, it is typical for exposure data to be summarized by calculating the geometric mean. Exposures tend to follow a lognormal distribution and the geometric mean is the value that represents the most "middle" value in such a distribution. However, if the concern of the end user is with total dose rather than with typical exposure levels, the arithmetic mean may be a more appropriate measure of central tendency, and should be provided with the assessment.

EXAMPLE

For the example chemical, several key issues were identified in the information supplied by the end users:

- Exposure of workers in the industry was of more interest than exposure of the general population.
- Worker exposure in the monomer industry was of more interest than worker exposure in the polymerization process. Worker exposure in handling of the finished polymer was of least interest.
- EPA was considering risk management options under TSCA. Since exposure may be limited to workers, a referral to OSHA was also possible. OSHA had no ongoing activities for the chemical at this time.
- Only inhalation exposure was of interest at this time.
- Only long-term exposure was of interest at this time.
- Specific descriptive statistics were requested.

Because the only data available were of Type 3, it was therefore necessary to conduct a monitoring program to obtain sufficient Type 1 data to conduct the types of analyses necessary to meet these needs.

STEP 4: IDENTIFY PARAMETERS AFFECTING EXPOSURE

Prior to statistical analysis, monitoring results must be classified into categories containing sufficient and reliable data so that meaningful analyses can be conducted (EPA, 87). The classification and organization of occupational exposure monitoring data are extremely important to the analysis and to the usefulness of the data for the end user. The classification and organization processes can be seen as the result of a compromise between two competing goals.

The first goal is to completely define the data set. If this were the single goal, the only data included would be those for which all parameters that can influence worker exposure were known, thus allowing definition of categories based on differences induced by all of these variables. For example, each category could be uniquely defined by process type, job title, worker activities, ambient control type (e.g., carbon adsorber), occupational control type (e.g., local exhaust ventilation), collection method, concentration of chemical in the process, demographics of the exposed worker, date the sample was taken, and any other parameter that could affect exposure or risk. The categories so defined would yield groups of exposure measurements (or groups of individual workers) expected to have the same or a similar exposure profile. Stated another way, the first goal is to define subsets of the data such that data within each subset are measuring the same thing, i.e., the subsets define homogeneous categories. Categories that are defined based on too few categorizing variables may lump together data that are not homogeneous.

The second goal, however, is to get categories with sufficient numbers of observations to allow meaningful statistical analyses. The power of any statistical analysis is greatly affected by sample size; large uncertainty can result when data sets are too small. The ability to make generalizations (extrapolations) is also limited when sample sizes are small. The number of observations within categories is inversely related to the number of categories (which is directly related to the number of parameters used to define the categories). Sample size is also reduced if observations have to be excluded from consideration because the values of variables potentially affecting those observations are missing or unknown.

The approach to balancing these two conflicting goals presented here has an industrial hygiene (qualitative) component and a statistical component. The industrial hygiene component is described in Step 4. The statistical component, described in Step 15, verifies the results of the industrial hygiene-based component and suggests possible re-categorization.

Thus, Step 4 consists of the critical process of identifying those parameters that are important in influencing worker exposure to the chemical under study. These exposure parameters will be used to define the categories (subsets or subpopulations) into which the exposure data will be classified.

CEB often develops categories of individuals with the same or similar exposure by first identifying the industrial process or unit operation during which exposure to the substance occurs, then identifying specific work activities or tasks associated with exposure, and identifying (or estimating) those workers associated with the activity or task, incorporating other information as appropriate. If monitoring data are available and job descriptions or job titles are given for the data, the engineer will need to evaluate whether

the job description or job title can be directly linked to a specific work activity or task. There are cases where the job title or description does reflect the work activity, but the converse is also true where job titles or job descriptions may be broader than the activities linked directly to the monitoring (Hawkins, 91).

If the job title is associated with a specific work activity, the engineer may determine that creating categories by industrial process/unit operation/job title/work activity/control type/etc. is appropriate. If the job title or description is not associated with a certain task or work activity, the engineer should try to obtain information on work activities associated with a personnel job title or description. If appropriate, an alternative is to make assumptions about the activities associated with the job title, based on knowledge of the process, professional judgment, etc. These assumptions should be fully documented and evaluated with other assumptions made during the assessment (see Step 5). It should also be noted that the identification of important exposure parameters is often refined as additional information is gathered during the exposure assessment.

Occupational control type is a variable that may affect worker exposure and which should often be considered when defining a classification scheme for exposure data.

The categories should also be designed with user needs in mind. This may include consideration of parameters that relate to risk assessment and regulatory considerations. All potential parameters will be used to create the preliminary exposure data matrix in Step 6.

A distinction may sometimes be made between exposure parameters that can be considered "explanatory" as opposed to those that are merely "blocking" factors. For example, it may be the case that exposures differ from one company to another, across plants, or with time. Although a statistical analysis may determine that plant-to-plant differences are significant, the factor, plant, does not "explain" why the exposures are different. Plant is not an explanatory parameter, it is what can be referred to as a blocking factor; the plant-to-plant differences may be present because of differences in occupational or ambient controls or other unknown factors that are directly related to exposure concentrations. Blocking factors are merely parameters within which exposures are expected to be similar. The factors that contribute to plant-to-plant differences, for example, may not be known or identified, and so it may sometimes be the case that such blocking variables need to be retained to account for differences in exposure levels. Nevertheless, the engineer is encouraged to identify explanatory parameters for the purposes of categorization. Retention of some blocking variables may be suggested, but their importance (as well as the importance of the proposed explanatory variables) will be tested statistically in Step 15.

The engineer should also consider the relative importance of the exposure factors considered for the classification. Based on his or her knowledge of the industry and the processes entailing exposure, he or she may be able to suggest that a small set of explanatory (and, perhaps, blocking factors) will be the most important for determining exposure. Parameters identified by the end user as important should be considered for the categorization, although, as discussed in Step 11, the expectations of the user may have to be modified in accordance with the availability of pertinent data. Job title, work practices, occupational controls, and production levels are typical examples of important parameters. One purpose of ranking the variables is to prioritize collection of additional information in these areas where necessary (see Steps 8 and 9).

Ideally, for risk assessment purposes, the exposure profiles for each exposed subpopulation defined by the parameters identified in this step should include the size of the group, the make-up of the group (age, sex, etc.), the source of the chemical, exposure pathways, the frequency and the intensity of exposure by each route (dermal, inhalation, etc.), the duration of exposure, and the form of the chemical when exposure occurs. Assumptions and uncertainties associated with each scenario and profile should be recorded and clearly discussed in the results presentation (EPA, 87).

The following parameters are presented as guidance to the CEB engineer as typical variables that can affect exposure and may be important in determining categories of similarly exposed individuals. They are presented in general order of their typical importance, but the actual importance of the parameter must be determined by the CEB engineer for the specific chemical and use.

- Type of sample - Sample type such as personal, area, ceiling, peak, etc. should be defined. In general, different sample types are not combined.

- Process type - Process should be defined by all characteristics that are likely to affect exposure. Examples include machine type (e.g., open-top vs. conveyorized degreaser), age of equipment, usage rate, and product (e.g., printing on paper vs. plastic).

- Job title - Job title is usually given with the monitoring data and may require combination of similar job descriptions (e.g., printer, letterpress operator, and press operator could be combined into a single category).

- Worker activities - Within a given job title, activities performed by the workers may vary in a significant way that can directly affect exposure.

- Worker location - The approximate location of the worker with respect to the source of the exposure is an important factor.

- Occupational control type (workplace practices) - Controls such as local exhaust ventilation (LEV) or general ventilation directly affect measured exposure. Other controls such such as respirators do not generally affect measured exposure but do affect actual worker exposure.

- Exposure period - The time period the worker is exposed to the chemical in a workday directly affects exposure. Frequency and duration of exposure are also important factors.

- Production levels - Exposure can relate directly to the volume of production at the facility.

- Operating frequency and duration - Total exposure relates directly to these variables.

- Concentration of chemicals in the process - The concentration of the chemical can directly affect the exposure of the workers. Such information is seldom available, however.

- Sampling strategy - The duration of the sampling and the sampling strategy can affect the accuracy of the measurements in characterizing the exposure.

- Ambient control type - Although such controls are installed primarily to reduce release of the chemical to the ambient air (e.g., refrigerated condenser, carbon adsorber, or baghouse), they may also increase or decrease occupational exposure.

- Company and location - Variables such as local regulations, differences between large and small companies, and regional differences in processes can affect worker exposure.

- Date of measurement - The date the measurement was taken can be indicative of the measurement method, the controls in use, and the effect of natural ventilation or other factors.

- Sample collection - Different collection methods, sampling times, validated range of the method, or method analytical techniques can affect the accuracy of the measurement and the detection limit.

- Source of data - Analysis by source of the data can help to identify potential biases in the data. Biases that are not evident in the review of data in Step 5 may be identified in Step 16.

- Demographics of the exposed worker - If health effects data show that a particular demographic group is susceptible (e.g., women of childbearing age), then whenever possible data should be categorized using this information. While this is not typically needed in an exposure assessment, it may be needed for a later health risk assessment.

- Industry - While four-digit SIC is preferable to two-digit SIC, OPPT assessments often focus on individual companies and/or facilities.

- Other - Depending on the process, controls implemented primarily for other substances may also reduce exposure to the substance of concern (e.g., LEV at the raw material transfer operation).

EXAMPLE

For the example data set, the following were identified as potentially important parameters:

- Sample type
- Job title
- Process type
- Occupational control
- Company
- Sample collection method
- Industry

While data were collected for other parameters discussed in this section, emphasis was placed on verifying information on these seven parameters. Note that the "blocking" variables, company and industry have been retained. Industry, in particular was retained because the end user had specified that the monomer industry needed to be considered separately from the polymer industry.

STEP 5: IDENTIFY UNCERTAINTIES, ASSUMPTIONS, AND BIASES

Uncertainties and assumptions are identified and recorded to allow their clear recognition by the end user. This step initiates that process. All data should be examined for any characteristics that may represent a nonrandom selection process or a systematic error (bias) in sampling or analysis. It may be helpful to review the list of important parameters to assist in identifying uncertainties, assumptions, and biases. All important uncertainties, assumptions, and biases are identified, and for purposes of grouping like exposure, these should be as specific as possible. In preparing the risk assessment, more general information on uncertainties, assumptions, and biases may be acceptable. Uncertainties, assumptions, and biases will be evaluated in Step 18 to determine any influence on estimates of worker exposure in one or more groups. Steps 5 and 18 are extremely important but may be difficult to execute.

A. Uncertainties

Examples of problems that give rise to typical uncertainties in the input and output of an exposure analysis include:

- Data manipulation errors either by the persons collecting the monitoring data or during the analysis.
- The inherent uncertainty in a small data set (e.g., day-to-day and worker-to-worker variability are not accounted for).
- Uncertainties regarding differences in chemical concentration, throughput, or other process related variables.
- Use of an unknown monitoring or analysis method.
- Assumptions made from secondary sources that were applied to the primary data.
- Uncertainties of values below the detection limit.
- Possible interference of other chemicals with a specific test method.
- Uncertainty regarding missing or incomplete information needed to fully define the exposure.
- The use of generic or surrogate data when site-specific data are not available.
- Errors in professional judgment.

In evaluating and reporting uncertainty associated with measurements, the three most important categories of errors are sampling errors, laboratory analysis errors, and data manipulation errors (EPA, 92). There are two kinds of sampling errors: systematic errors (often referred to as biases) that result from the sampling process, and random errors that result from the variability of both the population and the sampling process. While random error cannot be eliminated, its effects can be minimized by using sampling strategies and by having sufficiently large data sets. Systematic errors can result from faulty calibration of critical components such as flow meters, thermometers, pressure sensors, sieves, or other sampling devices.

Other systematic errors can result from contamination, losses, interactions with containers, deteriorations, or displacement of phase or chemical equilibria (EPA, 92).

Generally, laboratory errors are smaller than sampling errors. Calibration is a major source of systematic error in analysis. Other sources of error include chemical operations such as sample dissolution, concentration, extraction, and reactions (EPA, 92).

Data manipulation errors include errors of calculation, errors of transposition, errors of transmission, use of wrong units, use of improper conversion factors, spatial or temporal averaging information loss, and misassociation errors that confuse samples and numerical results.

B. Assumptions

Throughout the analysis, assumptions must be made about the data. Many assumptions are made in response to uncertainties identified in the data. These assumptions must be clearly listed and their effect on the results quantified if possible. Examples of typical assumptions that are made during exposure analysis include:

- That plants and workers were randomly selected and that they represent the industry as a whole. (It should be noted that this is almost never true; if it is known not to be true, this assumption should not be made.)
- That the controls in place when the data were collected represent typically maintained controls.
- That the value selected for use for a nondetected measurement accurately represents the actual exposure at those facilities.
- That estimates of ancillary information gathered from other sources also represent the facilities in the monitoring data set.
- That job activities performed during the exposure period represent typical activities for that job category.

- That estimates of the duration of tasks used to convert data to 8-hour TWA values are accurate.

C. Biases

Bias is a systematic error inherent in a method or caused by some feature of the measurement system (EPA, 92). Systematic errors in sample selection, sampling errors, laboratory analysis or data manipulation can cause the results to be biased. If the facilities and workers were not randomly selected and the selection process documented, then the data may also contain biases. Common features that may introduce bias include:

- Systematic sampling, laboratory, or data manipulation errors that have been identified.
- Selection of only "well-controlled" plants such as a NIOSH industry-wide survey conducted to identify good control technology.
- Selection of only large facilities.
- Large disparity between the number of samples at different facilities (e.g., OSHA vs. NIOSH data) could lead to bias, depending on how the data are weighted and whether there are underlying sampling biases.
- Data that represent only OSHA complaint visits.
- When sampling for compliance with a ceiling limit, sampling workers with the highest potential for exposure.
- Selection of only plants that are members of a trade association.
- Selection of only companies that voluntarily supplied monitoring data.
- Averaging of a measurement representing many workers with a measurement representing few workers.
- Use of sampling or analytical methods at concentrations for which they are not validated.
- Sampling strategy bias towards compliance sampling.

D. Development of Uncertainty/Assumptions List

In order to record and retain uncertainties, assumptions, and biases identified in the course of an occupational exposure assessment, a listing of the uncertainties and assumptions made at various steps will

be maintained. This list is initiated in this step and will initially contain uncertainties/assumptions associated with the data collection and classification. For example, in Step 4, some assumptions may have been required to relate job titles to specific activities. Moreover, there may have been uncertainties about the exposure profiles (number of workers, demographics of workers, source of chemical, etc.) for some of the groups defined by the important exposure parameters. These assumptions and uncertainties will be recorded in the uncertainty/assumption list.

In the course of following the guidelines defined in this document, other assumptions and uncertainties will be identified. All of them will be recorded on the uncertainty/assumption list for use in Step 18 (Treatment of Uncertainties, Assumptions, and Biases) and for presentation to the end-user with the quantitative results.

EXAMPLE

For the example chemical, a very detailed protocol and quality assurance plan were developed to select the facilities at which monitoring data would be collected. This protocol is more detailed than is typical but serves as an example of considerations that should be included to obtain a sample that is as representative as possible of the sample universe.

For manufacture of the example chemical monomer, the sample universe consisted of ten companies at 12 different plant locations. A walk-through survey was conducted at ten plants representing a 100 percent sample of the ten producers. The walk-through survey was used to gather information that was used to select a smaller sample set at which to conduct in-depth surveys. Monitoring data were collected at these in-depth surveys.

The purpose of the survey site selection strategy was to obtain a representative subset of monomer plants from which to characterize exposures by job title and work environment. To achieve this, the ten monomer production plants were divided into distinct subpopulations (strata) representing differences in the work place environment.

The strata were based on the presence or absence of three specific types of engineering controls, the mode of transportation (pipeline, rail car, tank truck, marine vessel) of the feed stock and product, and the existence of other production processes or final products at the plant. A single plant within each stratum was selected based on a scoring system that quantified the relative representativeness of each site. Four plants emerged as best representing the diversity of work environments seen in the example chemical monomer industry. In-depth surveys, including the collection of monitoring data, were conducted at these four facilities.

In the example data set, a serious potential bias in the analytical method for the chemical was identified. Potential interferences from C_4 chemicals made the measurements taken using previous methods suspect. Ways were investigated to mitigate the bias, but finally it was decided to exclude all data taken using the older analytical methods.

STEP 6: CREATE PRELIMINARY EXPOSURE DATA MATRIX

All data should be entered into a usable matrix using a personal computer for analysis. Software packages (spreadsheets, databases, etc.) are available with storage and retrieval capabilities that facilitate data analysis calculations. The matrix should be designed to be compatible with statistical programs that are likely to be used in the data analysis. Many statistical analysis packages have their own data matrix handling tools which provide a suitable, and in some cases preferable, alternative for data management. All parameters that were identified as having a potential impact on exposure, were requested by the end user, or were collected as ancillary information should be entered in the matrix. The use of a matrix will allow identification of missing information for some observations.

Inclusion of company name, plant location, and source of data in the data matrix is important because it provides a recordkeeping approach to allow easy referral of data back to the particular plant or study to obtain additional data. All potential variables should be entered into the data matrix and the field left blank when no data are found. Every effort should be made to fill in blanks in the matrix for all variables identified as important. An extra field or two should be included in the matrix for calculations such as converting to consistent units (Step 7). Also included would be any calculations made using assumptions such as the conversion of the TWA for the sampled time to an 8-hour TWA.

The exposure data matrix will be completed to the extent possible in Steps 7 through 9 by filling in missing information (where appropriate) and converting to consistent units. The revised exposure data matrix (Step 10) will serve to classify the data available and to assess the ability to meet the users' needs (Step 11). If possible, the data in the matrix will be used in the statistical analyses starting with Step 15.

EXAMPLE

Table 1 presents a partial example of the data matrix used in the example analysis. The full data set used in the analysis is presented in Appendix A. Only data on the important variables are presented in Table 1; however, data on all variables are included on the computer spreadsheet.

Table 1. Example Preliminary Exposure Data Matrix — Full Shift Personal Samples

Plant ID	Industry	ID0 (a)	Process Type	Job Title	Control Type (b)	Sample Duration (min)	8-hr TWA (ppm)	Control Description
M1	Monomer	1	Process area	Process technician	2	460	0.57	Single mechanical seals & open-loop bomb sampling
M1	Monomer	1	Process area	Process technician	2	470	≤0.18	Single mechanical seals & open-loop bomb sampling
M1	Monomer	1	Process area	Process technician	2	506	1.71	Single mechanical seals & open-loop bomb sampling
M1	Monomer	1	Process area	Process technician	2	457	0.74	Single mechanical seals & open-loop bomb sampling
M1	Monomer	1	Process area	Process technician	2	449	0.37	Single mechanical seals & open-loop bomb sampling
M2	Monomer	2	Process area	Process technician	1	260	≤0.27	Dual mechanical seals on pumps & closed-loop sampling
M2	Monomer	2	Process area	Process technician	1	437	0.70	Dual mechanical seals on pumps & closed-loop sampling
M2	Monomer	2	Process area	Process technician	1	452	1.23	Dual mechanical seals on pumps & closed-loop sampling
M2	Monomer	2	Process area	Process technician	1	452	2.37	Dual mechanical seals on pumps & closed-loop sampling
M2	Monomer	2	Process area	Process technician	1	454	2.98	Dual mechanical seals on pumps & closed-loop sampling
M2	Monomer	2	Process area	Process technician	1	451	0.46	Dual mechanical seals on pumps & closed-loop sampling
M3	Monomer	3	Process area	Process technician	2	432	4.19	Single mechanical seals & open-loop bomb sampling
M3	Monomer	3	Process area	Process technician	2	435	≤0.19	Single mechanical seals & open-loop bomb sampling
M3	Monomer	3	Process area	Process technician	2	437	1.34	Single mechanical seals & open-loop bomb sampling
M3	Monomer	3	Process area	Process technician	2	436	0.09	Single mechanical seals & open-loop bomb sampling
M4	Monomer	4	Process area	Process technician	2	461	1.76	Single mechanical seals & open-loop bomb sampling
M4	Monomer	4	Process area	Process technician	2	441	0.49	Single mechanical seals & open-loop bomb sampling
M4	Monomer	4	Process area	Process technician	2	459	2.11	Single mechanical seals & open-loop bomb sampling
M4	Monomer	4	Process area	Process technician	2	424	≤0.07	Single mechanical seals & open-loop bomb sampling
M4	Monomer	4	Process area	Process technician	2	473	1.00	Single mechanical seals & open-loop bomb sampling
M4	Monomer	4	Process area	Process technician	2	496	0.92	Single mechanical seals & open-loop bomb sampling
M4	Monomer	4	Process area	Process technician	2	434	≤0.08	Single mechanical seals & open-loop bomb sampling
M4	Monomer	4	Process area	Process technician	2	430	2.55	Single mechanical seals & open-loop bomb sampling
M4	Monomer	4	Process area	Process technician	2	456	0.29	Single mechanical seals & open-loop bomb sampling
M4	Monomer	4	Process area	Process technician	2	471	0.55	Single mechanical seals & open-loop bomb sampling
M4	Monomer	4	Process area	Process technician	2	465	0.27	Single mechanical seals & open-loop bomb sampling
M4	Monomer	4	Process area	Process technician	2	420	≤0.14	Single mechanical seals & open-loop bomb sampling

Table 1 (continued)

Plant ID	Industry	ID0 (a)	Process Type	Job Title	Control Type (b)	Sample Duration (min)	8-hr TWA (ppm)	Control Description
M1	Monomer	5	Control room	Process technician	1	466	≤0.02	General room ventilation
M1	Monomer	5	Control room	Process technician	1	447	≤0.02	General room ventilation
M1	Monomer	5	Control room	Process technician	1	485	≤0.02	General room ventilation
M2	Monomer	6	Control room	Process technician	1	455	0.25	General room ventilation
M2	Monomer	6	Control room	Process technician	1	451	≤0.08	General room ventilation
M2	Monomer	6	Control room	Process technician	1	452	0.52	General room ventilation
M3	Monomer	7	Control room	Process technician	1	442	≤0.04	General room ventilation
M3	Monomer	7	Control room	Process technician	1	425	≤0.11	General room ventilation
M4	Monomer	8	Control room	Process technician	1	453	1.87	General room ventilation
M4	Monomer	8	Control room	Process technician	1	449	1.70	General room ventilation
M2	Monomer	9	Loading area	Process technician	1	415	0.50	Slip-tube gauge
M2	Monomer	9	Loading area	Process technician	1	428	1.44	Slip-tube gauge
M2	Monomer	9	Loading area	Process technician	1	427	1.29	Slip-tube gauge
M3	Monomer	10	Loading area - railcar	Process technician	2	474	≤0.13	Magnetic gauge
M3	Monomer	10	Loading area - railcar	Process technician	2	260	≤0.12	Magnetic gauge
M3	Monomer	10	Loading area - railcar	Process technician	2	442	0.46	Magnetic gauge
M3	Monomer	10	Loading area/semi-tractor trailer	Process technician	2	484	2.40	Rotameter gauge
M3	Monomer	10	Loading area/semi-tractor trailer	Process technician	2	474	5.46	Rotameter gauge
M3	Monomer	10	Loading area/semi-tractor trailer	Process technician	2	446	0.08	Rotameter gauge
M4	Monomer	11	Loading area	Process technician	2	443	123.57	Slip-tube gauge
M4	Monomer	11	Loading area	Process technician	2	459	3.97	Slip-tube gauge

NOTE: Source of data: NIOSH/EPA. Laboratory analysis limit of detection ranged from 2 to 11 μg/sample, depending on the day of the analysis.

(a) IDO = Initial categories

(b) The following are the control types: 1) controlled, 2) uncontrolled, 3) laboratory with 12 air changes/hr, 4) 100% make-up air in laboratory, 5) 50% make-up air in laboratory, 6) 60% make-up air in laboratory.

STEP 7: CHECK FOR CONSISTENCY AND REASONABLENESS

Once the data have been loaded into the spreadsheet, the next step is to check them for consistency and reasonableness. It is recommended that, first, all the exposure measurements be converted to consistent units. This step describes some of the considerations related to conversion of units and the types of checks that can be made subsequently to verify that the results are reasonable.

For conversion of units, typically a standardized procedure consisting of grouping similar types of data, conversion to consistent concentration units, and conversion to consistent exposure periods can be used. For some data, however, all the information necessary to do the conversions is not known (e.g., actual exposure time period). In many of these cases, assumptions can be made that will allow use of the data in the analysis. All such assumptions should be recorded in the uncertainty/assumption list.

The general approach for conversion of data into consistent units is the following:

- Grouping of like types of data (e.g., 15 minute, long term, area, personal),
- Conversion to consistent concentration units (e.g., mg/m^3 or ppm),
- Conversion to consistent exposure periods when defensible (e.g., 8-hour TWA), and
- Estimation of missing information.

A. Grouping of Like Types of Data

It is extremely important that different types of samples not be averaged. For example, area samples generally do not represent personal exposure, and 15-minute peak and ceiling sampling should not be adjusted to represent full shift exposure. Specific data groupings that usually form like data sets and, as a general rule, should never be pooled into a single data set include:

- Area samples
- Personal samples
- Short term exposure estimates
- Long term exposure estimates

EXAMPLE

In the example data set only personal TWA samples will be used.

B. Conversion to Consistent Concentration Units

The end user should be consulted for guidance on preferable reporting units early in the project. Occupational exposure monitoring data are typically reported in either ppm or mg/m^3. NIOSH reports and journal articles report the occupational exposure values in either ppm or mg/m^3, while OSHA Inspection Summary Reports almost always report occupational exposure values in ppm. Before conducting statistical analysis on different data sets, all measurements need to be converted into similar units. Values in ppm can be converted to mg/m^3 by the following equation:

$$mg/m^3 = ppm \times \frac{MW}{24.45} \times \frac{P}{760} \times \frac{298}{(T + 273)}$$

where:

P	=	barometric pressure (mm Hg) of air sampled;
T	=	workplace temperature (°C) of air sampled;
24.45	=	molar volume (liter/g-mole) at 25°C and 760 mm Hg;
MW	=	molecular weight (g/g-mole);
760	=	standard pressure (mm Hg); and
298	=	standard temperature (°K).

EXAMPLE

Consider a case in which a chemical concentration is reported to be 5 ppm at a pressure of 760 mm Hg and 25°C. The molecular weight of the example chemical is 54.1 g/g-mole. The occupational exposure can then be converted from ppm by the following equation:

$$mg/m^3 = 5\ ppm \times \frac{54.1}{24.5} \times \frac{760}{760} \times \frac{298}{(25 + 273)}$$

$$mg/m^3 = 5\ ppm \times 2.213$$

Therefore, for the example chemical, a concentration of 5 ppm is equivalent to a concentration of 11.1 mg/m^3.

C. Conversion to Consistent Exposure Periods

NIOSH and OSHA exposure limits for chemicals are often based on 8-hour TWAs; therefore, occupational exposure monitoring data are often converted into 8-hour TWAs in order to compare worker exposures to these regulatory or recommended limits. Monitoring data collected from OSHA are typically reported as 8-hour TWAs because they are sampled for compliance with an 8-hour TWA Permissible Exposure Limit (PEL). OSHA TWA measurements may utilize a zero exposure for the unsampled portion of the 8-hour day when calculating the TWA. It may be useful to determine whether the sample represents an actual 8-hour sample or an 8-hour TWA. Some NIOSH reports and journal articles present data collected for less than an 8-hour time period. The measurement samples are literally only representative of the exposure period actually sampled. However, professional judgment or reliable knowledge may sometimes be used to extrapolate data collected for shorter time periods to an 8-hour TWA (Patty, 81). Where the exposure during the shorter period is representative of the exposure during the entire work period and the length of the work period is known, exposure values can be converted into 8-hour TWAs based on the shorter exposure duration.

Based upon the job description in the NIOSH report or journal article, an estimate of the number of worker hours per day related to each job category may be estimated. This should be done with caution as many times the sampling time was dictated by the analytical method or other cause not related to exposure and is not representative of the entire day. If the measurement sample is judged to be representative of the exposure period and the exposure period is less than 8 hours, then an exposure value not already reported as an 8-hour TWA can be adjusted to an 8-hour TWA as follows:

$$\text{8-hour TWA} = \text{exposed value} \times \frac{\text{exposed hours per day}}{8}$$

This approach is only valid when you can assume that there was no exposure during the remainder of the workday. This is a key assumption that should not be made without good information indicating that this is indeed the case.

Peak and ceiling measurements should never be converted to 8-hour TWA exposures. These measurements are best taken in a nonrandom fashion. That is, all available knowledge relating to the area, individual, and process being sampled are utilized to obtain samples during periods of maximum expected exposure (Patty, 81). Therefore these measurements by design are not representative of the longer work period. They are representative only of the time period over which they are taken, which usually corresponds to an applicable standard for peak or ceiling exposure.

EXAMPLE

While most samples were taken to represent 8-hour TWA exposures, some were not. Information gathered during the plant visit was used to estimate the exposure period for those measurements that did not represent 8-hour TWAs.

E. Checks for Consistency and Reasonableness

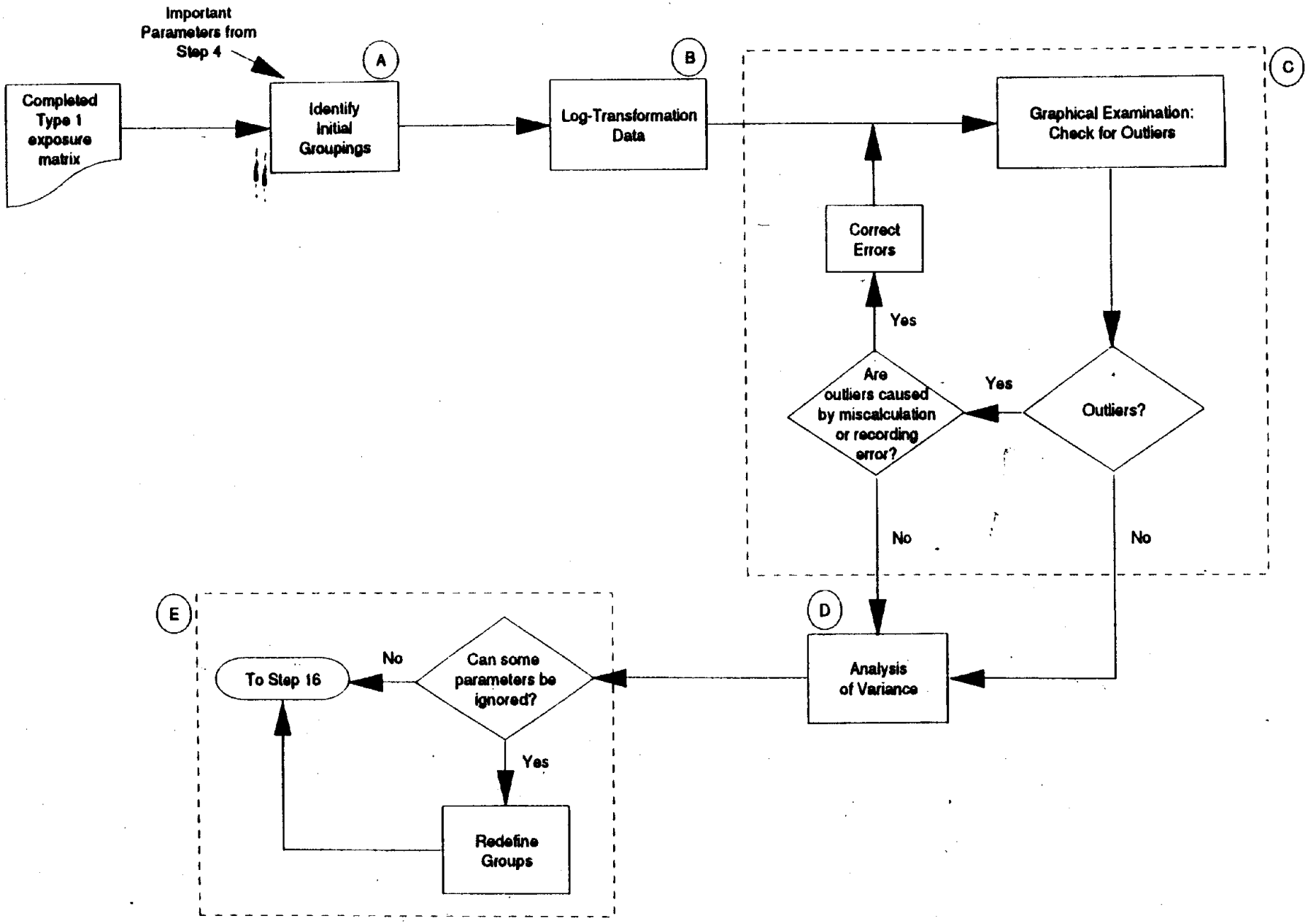

Figure 5. Flow Diagram for Step 15 (Define Groups for Analysis).

identified by number in Appendix A.

consult a statistician to help interpret problematic cases and to suggest supplemental analyses that may resolve the problems.

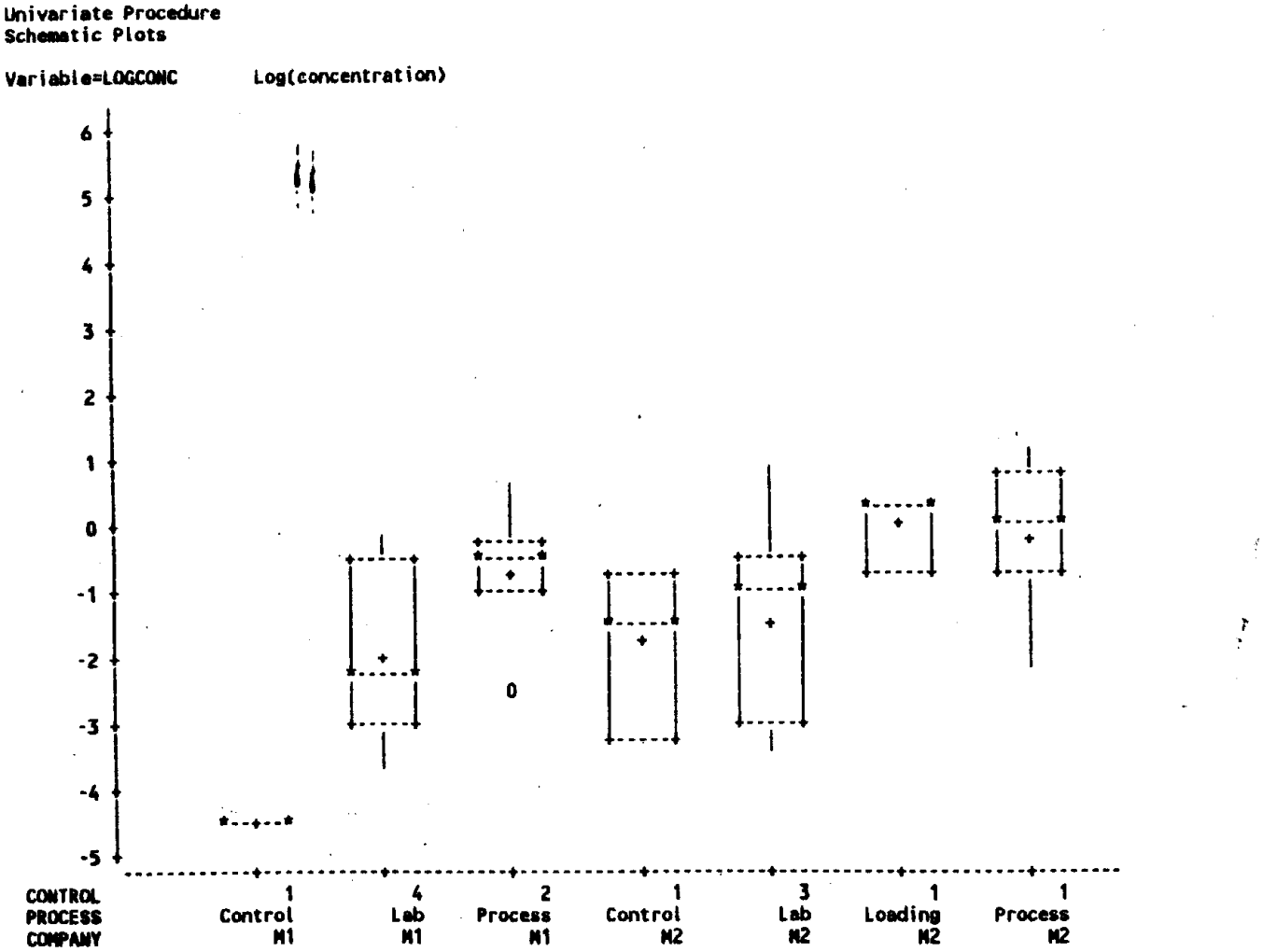

Figure 6: Box-and-Whisker Plot for Monomer Industry Categories

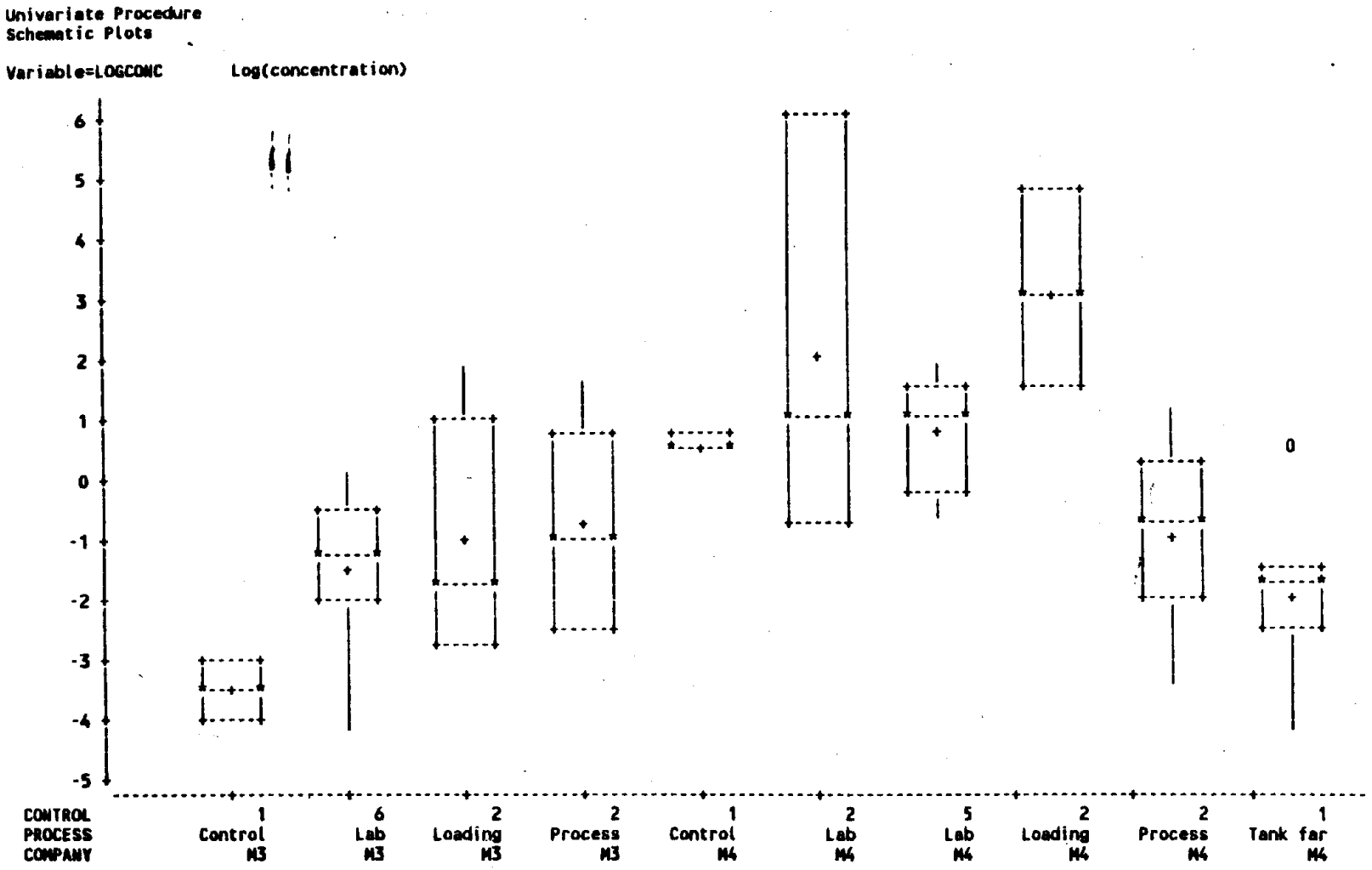

Figure 6: Box-and-Whisker Plot for Monomer Industry Categories

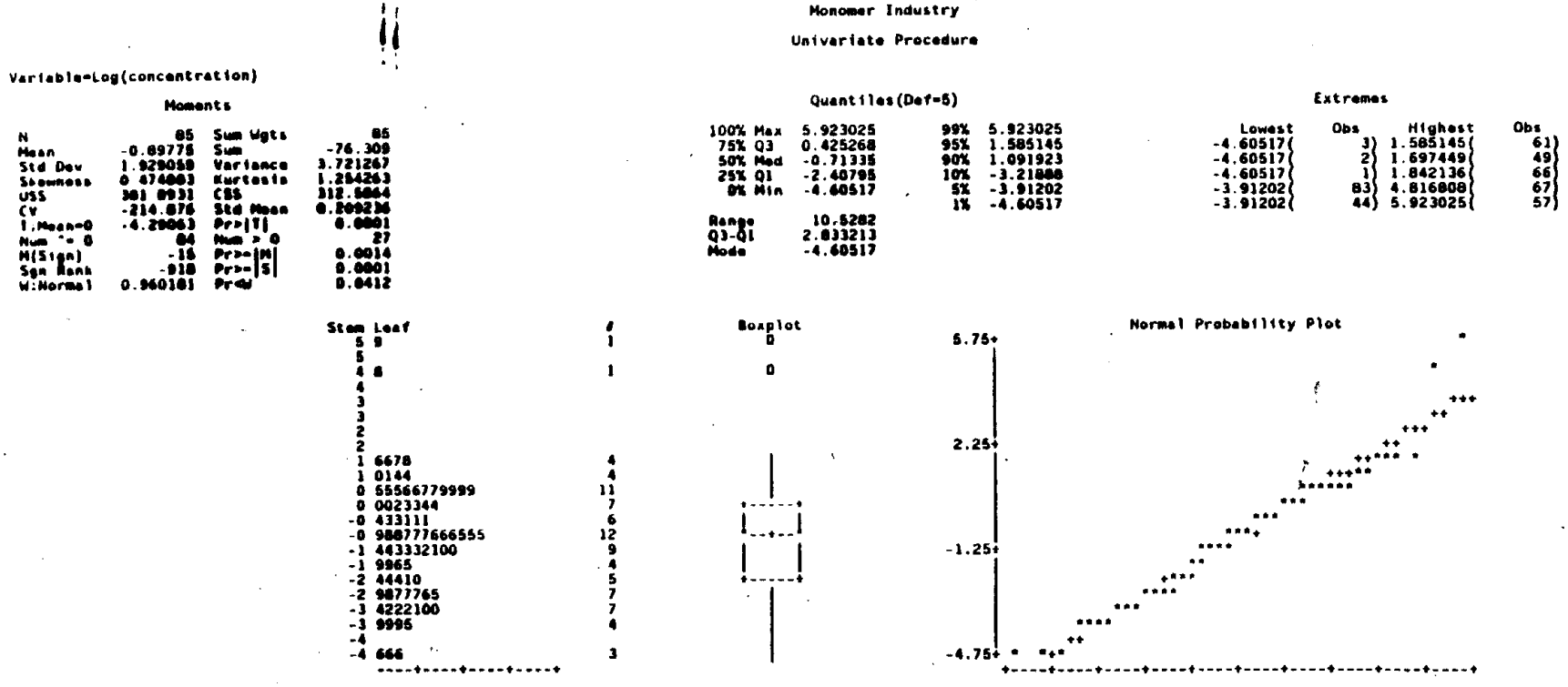

Figure 7. SAS Output for All Monomer Industry Categories Combined

General Linear Models Procedure
Class Level Information

Class	Levels	Values
COMPANY	4	M1 M2 M3 M4
PROCESS	5	Control room Lab Loading Process area Tank farm
CONTROL	6	1 2 3 4 5 6

Number of observations in data set = 85

Dependent Variable: LOGCONC Log(concentration)

Source	DF	Sum of Squares	Mean Square	F Value	Pr > F
Model	12	117.30527971	9.77543998	3.60	0.0003
Error	72	195.28116513	2.71223840		
Corrected Total	84	312.58644484			

R-Square	C.V.	Root MSE	LOGCONC Mean
0.375273	-183.4455	1.64688749	-0.89775311

Source	DF	Type III SS	Mean Square	F Value	Pr > F
COMPANY	3	19.11674454	6.37224818	2.35	0.0796
PROCESS	4	22.57606702	5.64401676	2.08	0.0922
CONTROL	5	18.94416819	3.78883364	1.40	0.2357

Figure 8: SAS Output for Test of Company, Process Type, and Control Type in Monomer Industry

General Linear Models Procedure
Class Level Information

Class	Levels	Values
PROCESS	5	Control room Lab Loading Process area Tank farm
CONTROL	6	1 2 3 4 5 6

Number of observations in data set = 85

Dependent Variable: LOGCONC Log(concentration)

Source	DF	Sum of Squares	Mean Square	F Value	Pr > F
Model	9	98.18853517	10.90983724	3.82	0.0005
Error	75	214.39790968	2.85863880		
Corrected Total	84	312.58644484			

R-Square	C.V.	Root MSE	LOGCONC Mean
0.314116	-188.3314	1.69075096	-0.89775311

Source	DF	Type III SS	Mean Square	F Value	Pr > F
PROCESS	4	49.48315238	12.37078810	4.33	0.0033
CONTROL	5	57.55457757	11.51091551	4.03	0.0027

Figure 9: SAS Output for Test of Process Type, and Control Type in Monomer Industry

General Linear Models Procedure
Class Level Information

Class	Levels	Values
COMPANY	5	P1 P2 P3 P4 P5
PROCESS	12	Control room Crumbing and dry Laboratory Maintenance Packaging Polymerization o Process area Purification Solutions and co Tank farm Unloading area Warehouse
CONTROL	2	1 2

Number of observations in data set = 431

Dependent Variable: LOGCONC Log(concentration)

Source	DF	Sum of Squares	Mean Square	F Value	Pr > F
Model	16	1279.36507409	79.96031713	47.59	0.0001
Error	414	695.61144185	1.68022087		
Corrected Total	430	1974.97651593			

R-Square	C.V.	Root MSE	LOGCONC Mean
0.647787	-47.86522	1.29623334	-2.70809000

Source	DF	Type III SS	Mean Square	F Value	Pr > F
COMPANY	4	487.37213944	121.84303486	72.52	0.0001
PROCESS	11	686.59871410	62.41806492	37.15	0.0001
CONTROL	1	0.96805416	0.96805416	0.58	0.4483

Figure 10 SAS Output for Test of Company, Process Type, and Control Type in Polymer Industry

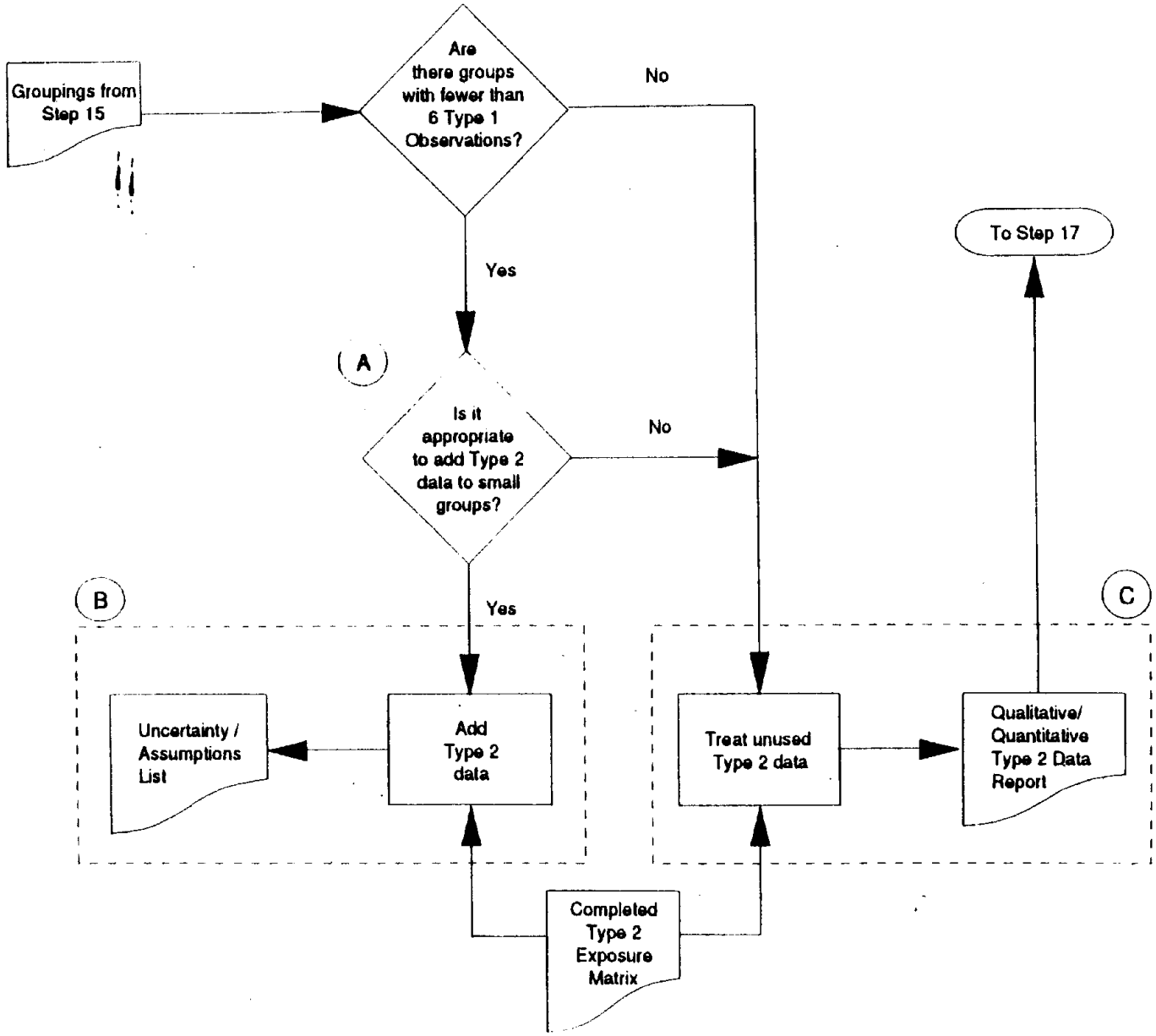

Figure 11. Flow Diagam for Step 16 (Treatment of Type 2 Data).

Table 3: Descriptive Statistics for Groups in Example Data Set

Group	No. of Samples	Descriptive Statistics Mean (ppm)	Std. Dev. (ppm)	Geom. Mean (ppm)	Geom. Std. Dev.
Monomer Control Room, Control 1	10	0.448	0.724	0.236	3.106
Monomer Lab, Control 2	3	2.610 *	--	--	--
Monomer Lab, Control 3	9	0.524	0.629	0.335	2.572
Monomer Lab, Control 4	6	0.298	0.357	0.191	2.569
Monomer Lab, Control 5	7	3.087	2.256	2.492	1.924
Monomer Lab, Control 6	7	0.350	0.304	0.264	2.116
Monomer Loading, Control 1	14	1.709	1.913	1.139	2.463
Monomer Loading, Control 2	8	17.010	43.110	6.243	4.120
Monomer Process Area, Control 1	6	1.312	1.131	0.994	2.107
Monomer Process Area, Control 2	21	0.918	1.054	0.603	2.502
Monomer Tank Farm, Control 1	5	0.160 *	--	--	--
P1, Crumbing and drying	9	0.043	0.019	0.040	1.515
P1, Lab	10	2.909	3.348	1.908	2.505
P1, Maintenance	34	0.857	2.310	0.298	4.277
P1, Packaging	30	0.039	0.031	0.031	2.003
P1, Polymerization or reaction	6	0.696	1.100	0.372	3.062
P1, Process area	6	0.118	0.122	0.082	2.346
P1, Purification	6	4.357	2.312	3.849	1.646
P1, Solutions and coagulation	9	0.027	0.008	0.026	1.343
P1, Tank farm	5	0.440 *	--	--	--
P1, Warehouse	2	0.020 *	--	--	--
P2, Control room	6	0.028	0.030	0.019	2.382
P2, Crumbing and drying	7	0.032	0.013	0.030	1.485
P2, Lab	14	0.636	1.267	0.285	3.547
P2, Maintenance	9	0.030	0.009	0.029	1.341
P2, Packaging	6	0.033	0.006	0.032	1.201
P2, Polymerization or reaction	29	0.077	0.144	0.036	3.417
P2, Solutions and coagulation	5	0.030 *	--	--	--
P2, Tank farm	3	0.360 *	--	--	--
P3, Lab	3	0.020 *	--	--	--
P3, Maintenance	4	0.020 *	--	--	--
P3, Polymerization or reaction	18	0.057	0.068	0.036	2.583
P3, Solutions and coagulation	4	0.020 *	--	--	--
P3, Tank farm	8	0.112	0.231	0.049	3.626
P3, Unloading area	2	14.600 *	--	--	--
P4, Crumbing and drying	13	0.016	0.020	0.010	2.682
P4, Lab	17	0.184	0.275	0.102	2.955
P4, Maintenance	7	0.004	0.004	0.003	2.140
P4, Packaging	20	0.006	0.006	0.004	2.374
P4, Polymerization or reaction	7	0.003	0.001	0.003	1.180
P4, Solutions and coagulation	3	0.003 *	--	--	--
P4, Tank farm	8	2.366	4.203	1.161	3.299
P4, Warehouse	11	0.004	0.002	0.004	1.627
P5, Crumbing and drying	6	0.055	0.031	0.048	1.697
P5, Lab	8	3.972	3.035	3.156	1.970
P5, Maintenance	16	1.200	1.253	0.830	2.360
P5, Packaging	23	0.058	0.034	0.050	1.730
P5, Polymerization or reaction	20	0.740	0.886	0.474	2.568
P5, Purification	12	9.523	6.727	7.778	1.889
P5, Solutions and coagulation	12	0.082	0.047	0.071	1.709
P5, Tank farm	6	3.020	1.750	2.613	1.713
P5, Warehouse	7	0.045	0.015	0.043	1.382

* Values marked by asterisks are medians for groups with less than 6 observations.

Table 4: Descriptive Statistics Presentation, Example Data Set

Group	No. of Exposed Workers	No. of Samples	Descriptive Statistics								Non-Detects	
			Minimum (ppm) (a)	Maximum (ppm) (a)	Median (ppm) (a)	Mean (ppm)	SE (ppm) (b)	Std. Dev. (ppm)	Geom. Mean (ppm)	Geom. Std. Dev.	No.	Percent
Monomer Control Room, Control 1	70	10	≤ 0.020	1.870	0.048	0.448	0.229	0.724	0.236	3.106	6	60
Monomer Lab, Control 2	93	3	0.420	373.540	2.610	--	--	--	--	--	0	0
Monomer Lab, Control 3	25	9	≤ 0.080	1.960	0.340	0.524	0.210	0.629	0.335	2.572	3	33
Monomer Lab, Control 4	40	6	0.030	0.870	0.110	0.298	0.146	0.357	0.191	2.569	0	0
Monomer Lab, Control 5	45	7	0.560	6.310	2.550	3.087	0.853	2.256	2.492	1.924	0	0
Monomer Lab, Control 6	61	7	≤ 0.040	0.890	0.280	0.350	0.115	0.304	0.264	2.116	1	14
Monomer Loading, Control 1	98	14	0.100	7.500	1.100	1.709	0.511	1.913	1.139	2.463	0	0
Monomer Loading, Control 2	106	8	≤ 0.080	123.570	1.430	17.010	15.242	43.110	6.243	4.120	2	25
Monomer Process Area, Control 1	111	6	≤ 0.270	2.980	0.960	1.312	0.462	1.131	0.994	2.107	1	17
Monomer Process Area, Control 2	95	21	≤ 0.070	4.190	0.550	0.918	0.230	1.054	0.603	2.502	5	24
Monomer Tank Farm, Control 1	83	5	≤ 0.040	1.530	0.155	--	--	--	--	--	3	60
P1, Crumbing and drying	166	9	0.014	0.071	0.040	0.043	0.006	0.019	0.040	1.515	0	0
P1, Lab	50	10	0.014	8.330	1.210	2.909	1.059	3.348	1.908	2.505	0	0
P1, Maintenance	110	34	0.014	11.020	0.100	0.857	0.396	2.310	0.298	4.277	0	0
P1, Packaging	36	30	0.012	0.154	0.028	0.039	0.006	0.031	0.031	2.003	0	0
P1, Polymerization or reaction	100	6	0.035	2.710	0.060	0.696	0.449	1.100	0.372	3.062	0	0
P1, Process area	80	6	≤ 0.006	0.304	0.075	0.118	0.050	0.122	0.082	2.346	1	17
P1, Purification	66	6	1.330	6.950	5.020	4.357	0.944	2.312	3.849	1.646	0	0
P1, Solutions and coagulation	260	9	0.019	0.046	0.025	0.027	0.003	0.008	0.026	1.343	0	0
P1, Tank farm	59	5	0.113	0.962	0.436	--	--	--	--	--	0	0
P1, Warehouse	10	2	0.014	0.020	0.017	--	--	--	--	--	0	0
P2, Control room	19	6	≤ 0.006	0.070	0.016	0.028	0.012	0.030	0.019	2.382	2	33
P2, Crumbing and drying	40	7	0.018	0.052	0.027	0.032	0.005	0.013	0.030	1.485	0	0
P2, Lab	63	14	0.029	4.120	0.044	0.636	0.339	1.267	0.285	3.547	0	0
P2, Maintenance	94	9	0.021	0.048	0.026	0.030	0.003	0.009	0.029	1.341	0	0
P2, Packaging	25	6	0.022	0.038	0.034	0.033	0.002	0.006	0.032	1.201	0	0
P2, Polymerization or reaction	105	29	≤ 0.008	0.780	0.033	0.077	0.027	0.144	0.036	3.417	2	7
P2, Solutions and coagulation	650	5	0.015	0.038	0.028	--	--	--	--	--	0	0
P2, Tank farm	59	3	0.123	0.436	0.362	--	--	--	--	--	0	0
P3, Lab	45	3	≤ 0.009	0.429	0.016	--	--	--	--	--	1	33
P3, Maintenance	74	4	0.011	0.026	0.020	--	--	--	--	--	0	0
P3, Polymerization or reaction	100	18	≤ 0.006	0.250	0.032	0.057	0.016	0.068	0.036	2.583	2	11
P3, Solutions and coagulation	460	4	≤ 0.006	0.164	0.019	--	--	--	--	--	1	25
P3, Tank farm	41	8	0.009	0.682	0.034	0.112	0.082	0.231	0.049	3.626	0	0
P3, Unloading area	45	2	0.770	28.510	14.640	--	--	--	--	--	0	0
P4, Crumbing and drying	24	13	≤ 0.005	0.081	0.013	0.016	0.006	0.020	0.010	2.682	4	31
P4, Lab	60	17	≤ 0.006	0.943	0.069	0.184	0.067	0.275	0.102	2.955	3	18
P4, Maintenance	61	7	≤ 0.006	0.013	0.003	0.004	0.001	0.004	0.003	2.140	6	86
P4, Packaging	400	20	≤ 0.006	0.026	0.003	0.006	0.001	0.006	0.004	2.374	16	80
P4, Polymerization or reaction	204	7	≤ 0.006	≤ 0.008	0.003	0.003	0.000	0.001	0.003	1.180	7	100
P4, Solutions and coagulation	315	3	≤ 0.005	≤ 0.008	0.003	--	--	--	--	--	3	100
P4, Tank farm	45	8	≤ 0.006	12.030	0.392	2.366	1.486	4.203	1.161	3.299	1	12
P4, Warehouse	56	11	≤ 0.005	≤ 0.018	0.003	0.004	0.001	0.002	0.004	1.627	10	91
P5, Crumbing and drying	39	6	0.033	0.116	0.043	0.055	0.013	0.031	0.048	1.697	0	0
P5, Lab	36	8	0.100	8.870	4.580	3.972	1.073	3.035	3.156	1.970	0	0
P5, Maintenance	80	16	0.072	3.890	0.655	1.200	0.313	1.253	0.830	2.360	0	0
P5, Packaging	44	23	≤ 0.014	0.144	0.042	0.058	0.007	0.034	0.050	1.730	1	4

Table 4: Descriptive Statistics Presentation, Example Data Set

Group	No. of Exposed Workers	No. of Samples	Descriptive Statistics							Non-Detects	
			Minimum (ppm) (a)	Maximum (ppm) (a)	Median (ppm) (a)	Mean (ppm)	SE (ppm) (b)	Std. Dev. (ppm)	Geom. Mean (ppm)	Geom. Std. Dev.	No.	Percent
P5, Polymerization or reaction	52	20	0.035	2.800	0.400	0.740	0.198	0.886	0.474	2.568	0	0
P5, Purification	90	12	2.770	24.140	7.580	9.523	1.942	6.727	7.778	1.889	0	0
P5, Solutions and coagulation	555	12	≤ 0.006	0.169	0.090	0.082	0.014	0.047	0.071	1.709	1	8
P5, Tank farm	41	6	1.070	6.010	2.760	3.020	0.714	1.750	2.613	1.713	0	0
P5, Warehouse	30	7	0.033	0.068	0.039	0.045	0.006	0.015	0.043	1.382	0	0

(a) The minimum, maximum, and median are provided as additional descriptive statistics.
(b) Standard error measures precision of the mean.

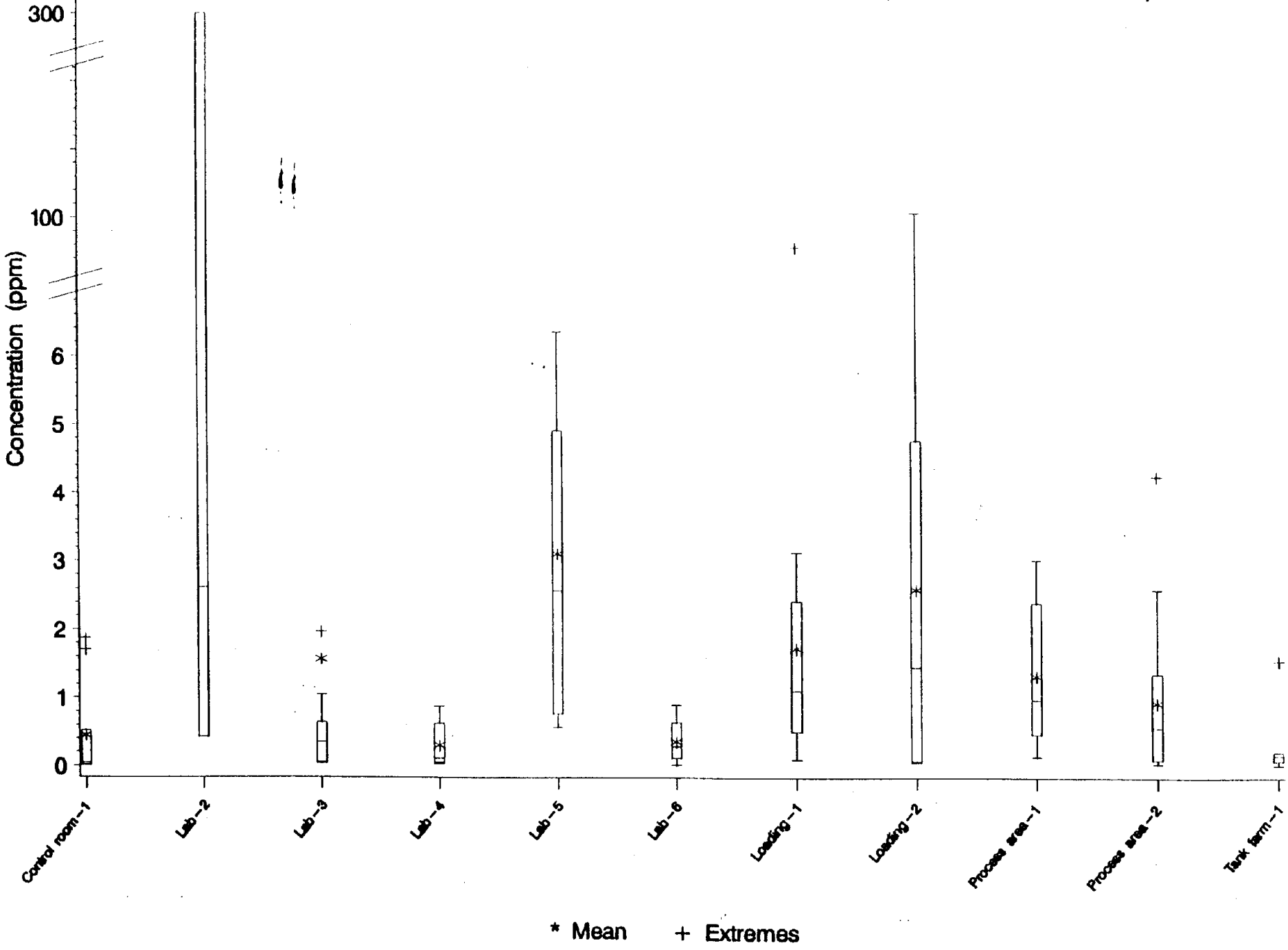

Figure 12. Box-and-Whisker Plot for Monomer Industry Groups

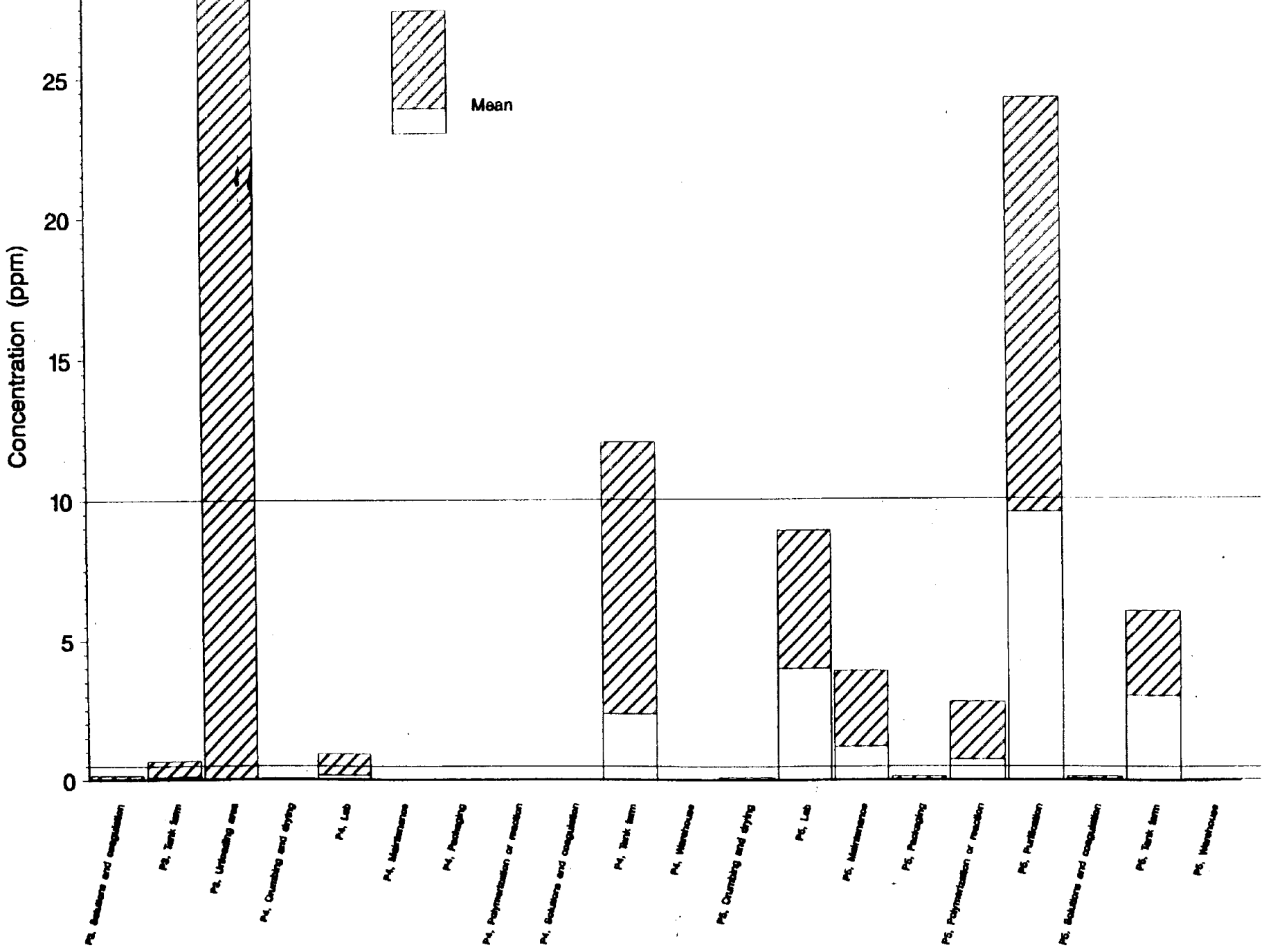

Figure 13. Example Bar Graph for Polymer Industry Groups: Means and Maxima Compared to 3 Target Levels

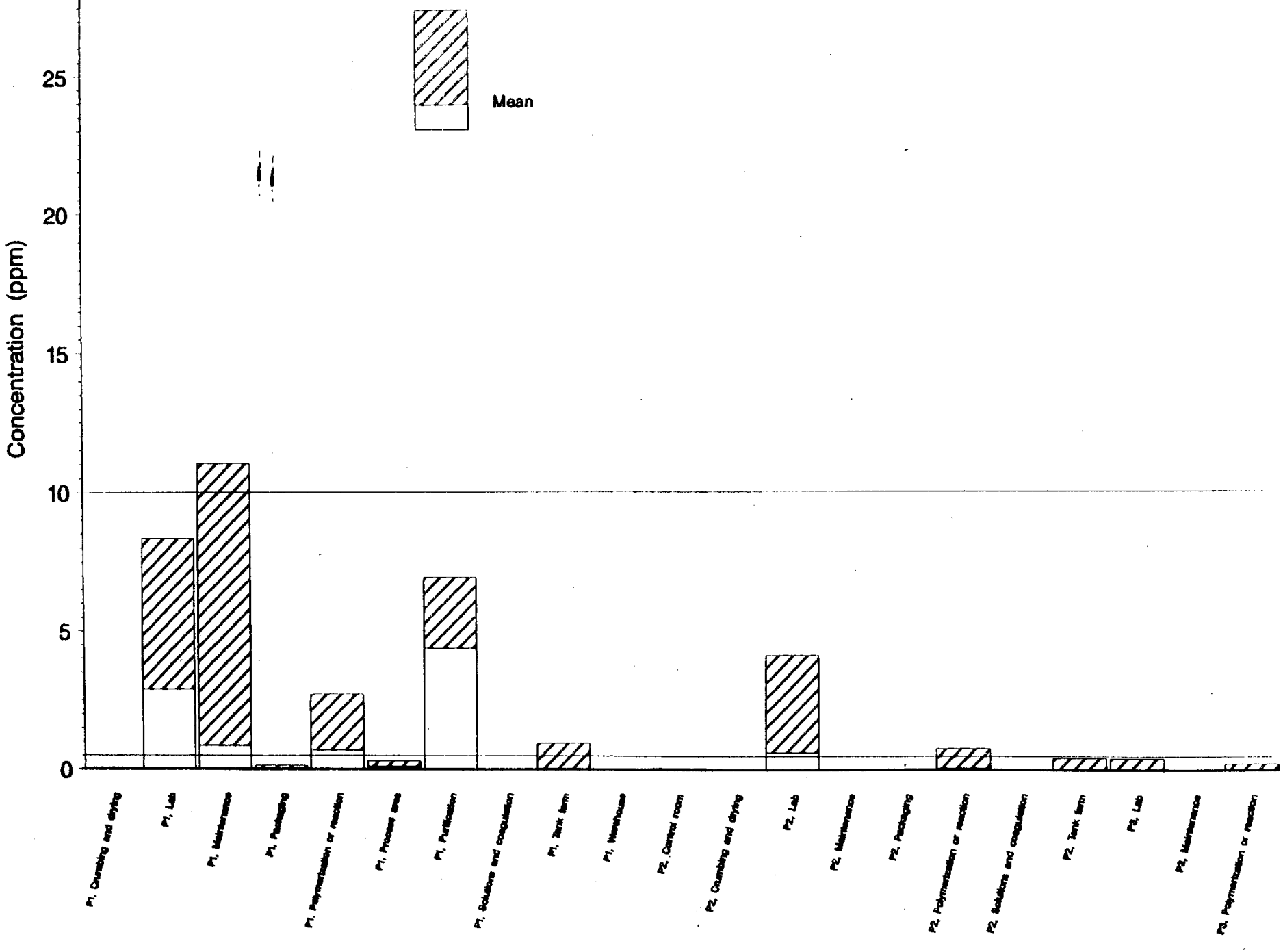

Figure 13. Example Bar Graph for Polymer Industry Groups: Means and Maxima Compared to 3 Target Levels

Uncertainty	Associated assumption	Reasonable possible variance of assumption.	Effect on results
For job category A the length of work day is not known for 30% of the monitoring data.	Length of work day assumed to be 6 hours.	Reasonable range is 5 to 7 hours.	Maximum 6% change in descriptive statistic for job category A (sensitivity analysis).
Actual exposure not known for values recorded as nondetected (5% of values).	A value of $L/\sqrt{2}$ was assumed. L = 1 ppm. ND = 0.71 ppm.	A value of L/2 could better represent actual exposure.	Maximum 2% change in overall descriptive statistic (sensitivity analysis).
NIOSH indicates that data for industry B represents "well controlled" facilities.	None made.	NIOSH personnel roughly estimated that exposures at well controlled facilities can be 20% lower than the industry average.	Descriptive statistics for industry B may underestimate exposure by up to 20% (NIOSH estimate).
Plants in the industry C data set were not randomly selected but rather all available data was used.	The data set for industry C represents the industry as a whole.	Not quantifiable.	Unknown
For job category D only OSHA compliance data were used.	None made.	Not quantifiable.	Facilities where OSHA complaints are made may have higher exposure than the industry as a whole (engineering judgment).
etc.	etc.	etc.	etc.

Figure 14. Example Format for Presentation of Assumptions and Uncertainties.

REFERENCES AND BIBLIOGRAPHY

(Aitchison, 57) Aitchison, J. and J.A.C. Brown. The Lognormal Distribution. Cambridge University Press. London. 1957.

(Armstrong, 92) Armstrong, Ben C. Confidence Intervals for Arithmetic Means of Lognormally Distributed Exposures. American Industrial Hygiene Association Journal. 53:481-485. 1992.

(Attfield, 92) Attfield, M.D. and P. Hewett. Exact Expressions for the Bias and Variance of Estimators of the Mean of a Lognormal Distribution. American Industrial Hygiene Association Journal. 53:432-435. 1992.

(Bickel, 77) Bickel, P.J. and K.A. Doksum. Mathematical Statistics. Holden-Day, Ind. San Francisco, CA. 1977.

(BMDP/PC) BMDP Statistical Software, Inc., 1440 Sepulveda Blvd., Suite 316, Los Angeles, CA 90025. 213/479-7799.

(Box, 78) Box, George E., William G. Hunter, and J. Stuart Hunter. Statistics for Experimenters. John Wiley & Sons. New York, New York. 1978.

(Buringh, 91) Bruingh, Eltjo and Roel Lanting. Exposure Variability in the Workplace: Its Implications for the Assessment of Compliance. American Industrial Hygiene Association Journal 52:6-13. 1991.

(CMA, 86) Chemical Manufacturers Association. Papers Presented at the Workshop on Strategies for Measuring Exposure. December 9 and 10, 1986.

(Cochran, 63) Cochran, William G. Sampling Techniques. John Wiley and Sons, Inc. New York, New York. 1963.

(Cohen, 61) Cohen, C.A. Tables for Maximum Likelihood Estimates: Singly Truncated or Singly Censored Samples. Technometrics 3:535. 1961.

(Cohen, 78) Cohen, Clifford, et al. Statistical Analysis of Radionuclide Levels in Food Commodities, prepared for U.S. Food and Drug Administration, Washington, D.C. September 15, 1978.

(Conover, 80) Conover W.J. Practical Nonparametric Statistics. 2nd ed. John Wiley & Sons. New York, New York. 1980.

(Corn, 79) Corn, Morton and Nurtan A. Esmen. Workplace Exposure Zones for Classification of Employee Exposures to Physical and Chemical Agents. American Industrial Hygiene Association Journal. 40:47-57. 1979.

(Cox, 81) Cox, David C. and Paul Baybutt. Methods for Uncertainty Analysis: A Comparative Survey. Risk Analysis. 1:251-258. 1981.

(Crump, 78) Crump, Kenny S. Estimation of Mean Pesticide Concentrations When Observations are Detected Below the Quantification Limit. Prepared for the Food and Drug Administration. Washington, D.C. April 9, 1978.

(Damiano, 86) Damiano, Joe. The Alcoa Sampling and Evaluation Guidelines, presented at the Workshop on Strategies for Measuring Exposure (CMA). December 9 and 10, 1986.

(Damiano, 89) Damiano, Joe. A Guideline for Managing the Industrial Hygiene Sampling Function. American Industrial Hygiene Association Journal. July 1989.

(Daniel, 78) Daniel, Wayne S. Applied Nonparametric Statistics. Houghton Mifflin Company. Boston, Massachusetts. 1978.

(Devore, 82) Devore, Jay L. Probability and Statistics for Engineering and the Sciences. Table A.3 Standard Normal Curve Areas. p. 620. Brooks/Cole Publishing Company, Monterey, CA. 1982.

(Dixon, 83) Dixon, S.W., et al. E.I. du Pont de Nemours. Management of Air Sampling Results, presented at the American Industrial Hygiene Conference, Philadelphia, Pennsylvania. May 25, 1983.

(Eisenhart, 68) Eisenhart, Churchill. Expression of the Uncertainties of Final Results. Science. pp 1201-1204. June 1968.

(EPA, 78) Environmental Protection Agency. Source Assessment: Analysis of Uncertainty - Principles and Application EPA/600/13. Industrial Environmental Research Laboratory, Research Triangle Park, North Carolina. August 1978.

(EPA, 87) Environmental Protection Agency. The Risk Assessment Guidelines of 1986. EPA/600/8-87/045. Washington, D.C. August 1987.

(EPA, 92) Environmental Protection Agency. Exposure Assessment Guidelines. EPA/600/Z-92/001. Washington, D.C. 1992.

(Esmen, 77) Esmen, Nurtan A. and Yehia Y. Hammad. Log-Normality of Environmental Sampling Data. Journal of Environmental Science and Health, A12 (1 & 2), pp. 29-41. 1977.

(Hansen, 83) Hansen, Morris H., et al. An Evaluation of Model-Dependent and Probability-Sampling Inferences in Sample Surveys. Journal of the American Statistical Association. pp. 776-793. December 1983.

(Hawkins, 92) Hawkins, Neil C., Michael A. Jayjock, and Jeremiah Lynch. A Rationale and Framework for Establishing the Quality of Human Exposure Assessments. American Industrial Hygiene Association Journal 53:34-41. 1992.

(Hawkins, 91) Hawkins, Neil C., et al. A Strategy for Occupational Exposure Assessments. American Industrial Hygiene Association. Akron, Ohio. 1991.

(Hoaglin, 83) Hoaglin, David D., et al. Understanding Robust and Exploratory Data Analysis. John Wiley and Sons, Inc. New York, New York. 1983.

(Hornung, 90) Hornung, Richard W., and Lawrence D. Reed. Estimation of Average Concentration in the Presence of Nondetectable Values. Applied Occup Environ Hyg 5(1):46-51. 1990.

(Hornung, 91) Hornung, Richard W. Statistical Evaluation of Exposure Strategies. Applied Occup Environ Hyg 6(6):516-520. 1991.

(IT, 91) IT Environmental Programs, Inc. Preparation of Engineering Assessments, Volume I: CEB Engineering Manual. U.S. Environmental Protection Agency/Office of Toxic Substances. Washington, D.C. Contract No. 68-D8-0112. February, 1991.

(Jackson, 85) Jackson, R.A., and A. Behar. Noise Exposure - Sample Size and Confidence Limit Calculation. American Industrial Hygiene Association Journal. 46:387-390. 1985.

(Johnson, 70) Johnson Norman L. and Samuel Kotz. Continuous Univariate Distributions-1. John Wiley & Sons. New York, New York. 1970.

(Karch, 88) Karch, Nathan J. Testimony of Nathan J. Karch, Ph.D. on the Quantitative Risk Assessments Included in the Proposed Rule on Air Contaminants of the Occupational Safety and Health Administration. July 28, 1988.

(Koek, 88) Koek, Karin E., at al. Encyclopedia of Associations 1989, 3 vol. Gale Research, Inc. Detroit, Michigan. 1988.

(Koizumi, 80) Koizumi, Akio, et al. Evaluation of the Time Weighted Average of Air Contaminants with Special References to Concentration Fluctuation and Biological Half Time. American Industrial Hygiene Association Journal. pp 693-699. October 1980.

(Lee, Undated) Lee, Shin Tao, et al. A Calculation and Auto Selection of Simple Statistical Analyses for Industrial Hygiene Data, NIOSH, Cincinnati, Ohio. Undated.

(Lemasters, 85) Lemasters, Grace K., Arch Carson, and Steven J. Samuels. Occupational Styrene Exposure for Twelve Product Categories in the Reinforced-Plastics Industry. American Industrial Association Journal 46:434-441. 1985.

(Lilliefors, 67) Lilliefors, H.W. On the Kolmogorov-Smirnov test for normality with mean and variance unknown. Journal of the American Statistical Association 62:399-402. 1967.

(Massey, 51) Massey, Frank J., Jr. The Kolmogorov-Smirnov Test for Goodness of Fit. Journal of the American Statistical Association. pp. 68-78. 1951.

(McBride, 91) McBride, Judith B. A Study of Spatial Correlation in a Workplace Exposure Zone. Presented at American Industrial Hygiene Association Conference and Exposition, Salt Lake City, Utah. May 23, 1991.

(Nicas, 91) Nicas, Mark, Barton P. Simmons, and Robert C. Spear. Environmental versus analytical Variability in Exposure Measurements. American Industrial Hygiene Association Journal 52:553-557. 1991.

(NIOSH, 77) National Institute for Occupational Safety and Health. Occupational Exposure Sampling Strategy Manual. DHEW (NIOSH) Publication No. 77-173. Cincinnati, Ohio. 1977.

(NIOSH, 84) National Institute for Occupational Safety and Health. NIOSH Manual of Analytical Methods. Third Edition. Cincinnati, Ohio 1984.

(Oldham, 65) Oldham, P.D. On Estimating the Arithmetic Means of Lognormally-Distributed Populations. Biometrics. 213:235-239. 1965.

(Olsen, 91) Olsen, Erik, Bjarne Laursen, and Peter S. Vinzents. Bias and Random Errors in Historical Data of Exposure to Organic Solvents. American Industrial Hygiene Association Journal 52:204-211. 1991.

(OSHA, 90) Occupational Safety and Health Administration. OSHA Technical Manual. Issued by OSHA February 5, 1990.

(OSHA, 85) OSHA Chemical Information File. OSHA Directorate of Technical Support. June 13, 1985.

(OSHA, Unpublished) OSHA Manual of Analytical Methods (Unpublished).

(Patty, 81) Patty, Frank A. Patty's Industrial Hygiene and Toxicology, 3rd Edition. Volumes 1 through 3 General Principles, Statistical Design and Data Analysis Requirements. John Wiley & Sons, New York, New York, 1981.

(Powell, 88) Powell, R.W. A Method for Calculating a Mean of a Lognormal Distribution of Exposures. Exxon Research and Engineering Company. Florham Park, New Jersey. 1988.

(Preat, 87) Preat, Bernard. Application of Geostatistical Methods for Estimation of the Dispersion Variance of Occupational Exposures. American Industrial Hygiene Association Journal. 48:877-884. 1987.

(Rappaport, 91) Rappaport, S.M. Assessment of Long-term Exposures to Toxic Substances in Air. Annuals of Occupational Hygiene 35:61-121. 1991.

(Rappaport, 87) Rappaport, S.M. and S. Selvin. A Method for Evaluating the Mean Exposure from a Lognormal Distribution. American Industrial Hygiene Association Journal. 48:374-379. 1987.

(Re, 85) Re, M. Microcomputer Programs for the Evaluation of Predictable Long-Term Exposure. American Industrial Hygiene Association Journal. 46:369-372. 1985.

(Rock, 82) Rock, James C. A Comparison Between OSHA-Compliance Criteria and Action-Level Decision Criteria. American Industrial Hygiene Association Journal. 43:297-313. 1982.

(Samuels, 85) Samuels, Steven J., Grace K. Lemasters, and Arch Carson. Statistical Methods for Describing Occupational Exposure Measurements. American Industrial Hygiene Association Journal. 46:427-433. 1985.

(Searle, 92) Searle, Shayle R., George Casella, and Charles E. McCulloch. Variance Components. John Wiley & Sons. New York, New York. 1992.

(Schneider, 91) Schneider, Thomas, Ib Olsen, Ole Jorgensen, and Bjarne Lauersen. Evaluation of Exposure Information. Applied Occupational and Environmental Hygiene 6:475-481. 1991.

(Selvin, 91) Selvin, S. Review of draft version of "Guidelines for Statistical Analysis of Occupational Exposure Data." Submitted to EPA Chemical Engineering Branch. November, 1991.

(Selvin, 89) Selvin, S. and S.M. Rappaport. A Note on the Estimation of the Mean Value from a Lognormal Distribution. American Industrial Hygiene Association Journal. 50:627-630. 1989.

(Selvin, 87) Selvin, S., et al. A Note on the Assessment of Exposure Using One-Sided Tolerance Limits. American Industrial Hygiene Association Journal. 48:89-93. 1987.

(Sokal, 81) Sokal, Robert R. and James Rohlf. Biometry. W. H. Freeman and Company. New York, New York. 1981.

(SPSS/PC) SPSS, Inc., 444 N. Michigan Ave., Suite 3000, Chicago, IL 60611. 312/329-2400.

(Stoline, 91) Stoline, Michael R. An Examination of the Lognormal and Box and Cox Family of Transformations in Fitting Environmental Data. Environmetrics 2:85-106. 1991.

(SYSTAT) Systat, Inc., 1800 Sherman Ave., Evanston, IL 60201. 312/864-5670.

(Tiat, 92) Tiat, Keith. The Workplace Exposure Assessment Expert System (WORKSPERT). American Industrial Hygiene Association Journal 53:84-98. 1992.

(Tuggle, 81) Tuggle, R. M. The NIOSH Decision Scheme. American Industrial Hygiene Association Journal. 42:493-498. 1981.

(Tuggle, 82) Tuggle, R. M. Assessment of Occupational Exposure Using One-Sided Tolerance Limits. American Industrial Hygiene Association Journal. 43:338-346. 1982.

(Waters, 91) Waters, Martha A., Steve Selvin, Stephen M. Rappaport. A Measure of Goodness-of-Fit for the Lognormal Model Applied to Occupational Exposures. American Industrial Hygiene Association Journal 52:493-502. 1991.

(Waters, 90) Waters, Martha A. Some Statistical Considerations in Chemical Exposure Assessment. Ph.D. dissertation, University of California at Berkeley, California. 1990.

(Whitmore, 85) Whitmore, Roy W. Methodology for Characterization of Uncertainty in Exposure Assessments. U.S. EPA, Office of Health and Environmental Assessment, EPA/600/8-85/009. August 1985.

(Woodruff, 71) Woodruff, Ralph S. A Simple Method for Approximating the Variance of a Complicated Estimate. Journal of the American Statistical Association. pp. 411-414. June 1971.

GLOSSARY OF TERMS

Accuracy - the measure of the correctness of the data, as given by the difference between the measured value and the true value.

Sample Mean - the sum of all the measurements in the data set divided by the number of measurements in the data set.

Bias - a systematic error inherent in a method or caused by some feature of the measurement system.

Bimodal Distribution - a probability density function with two relative maxima values.

Bounding Estimate - an estimate of exposure that is higher than the exposure of the individual in the population with the highest exposure. Bounding estimates are useful in constructing statements such as ".. exposure is not greater than" the estimated value.

Confidence Interval - a range of values that contains the true value of a parameter in a distribution a predetermined proportion of time if the process of determining the value is repeated a number of times.

Descriptive Statistics - statistics that describe conditions and events in terms of the observed data; use is made of tables, graphs, ratios, and typical parameters such as location statistics (e.g., arithmetic mean) and dispersion statistics (e.g., variance).

Frequency Histogram - a graphical representation of a frequency distribution, typically using bars to exhibit the frequency or relative frequency of occurrence of each value or group of values in a data set.

Geometric Mean - the n^{th} root of the product of n values.

High End Estimate - a plausible estimate of individual exposure for those persons at the upper end of an exposure distribution, conceptually above the 90^{th} percentile, but not higher than the individual in the population with the highest exposure.

Homogeneous Categories - groups or categories with the same or similar modifying attributes.

Limit of Detection - the minimum concentration of an analyte that, in a given matrix and with a specific method, has a 99% probability of being identified, qualitatively or quantitatively measured, and reported to be greater than zero.

Log-normal Distribution - a probability distribution restricted to positive real values. If the random variable Y has a log-normal distribution, then $X = \log_e Y$, then X has a normal distribution.

Maximally Exposed Individual (MEI) - a semiquantitative term referring to the extreme uppermost portion of the distribution of exposures. For consistency, this term should refer to the portion of the individual exposure distribution that conceptually falls above the 98^{th} percentile of the distribution, but is not higher than the individual with the highest exposure.

Maximum-Likelihood Estimate - an estimate based on finding the values of parameters that give the maximum value of the likelihood function. The likelihood function is the probability of observing the data, as a function of the parameters defining a distribution. The maximum likelihood approach is applicable whenever the underlying distribution of the data is known or assumed. It is a common statistical estimation procedure.

Median - the value in a measurement data set such that half the measured values are greater and half are less.

Nonparametric Statistical Methods - methods that do not assume a functional form with identifiable parameters for the statistical distribution of interest (distribution-free methods).

Normal Distribution - a symmetric probability distribution whose maximum height is at the mean, applicable to positive and negative real numbers. The normal distribution is the common "bell-shaped" curve. Also called a Gaussian distribution.

Precision - a measure of the reproducibility of a measured value under a given set of conditions.

Probability Sampling - sampling method in which each population element has a known and nonzero probability of being selected. Basic probability sampling methods include simple random sampling, stratified sampling, and cluster sampling.

Quantification Limit - the concentration of analyte in a specific matrix for which the probability of producing analytical values above the method detection limit is 99%.

Random Sampling - the selection of a sample of size n in such a way that each possible sample of size n has the same chance of being selected.

Reasonable Worst Case - a semiquantitative term referring to the lower portion of the high end of the exposure distribution. For consistency, it should refer to a range that can conceptually be described as above the 90^{th} percentile in the distribution, but below about the 98^{th} percentile.

Representativeness - the degree to which a sample is, or samples are, characteristic of the whole medium, exposure, or dose for which the samples are being used to make inferences.

Sample - a small part of something designed to show the nature or quality of the whole. Exposure-related measurements may be samples of exposures of a small subset of a population for a short time, for the purpose of inferring the nature and quality of the parameters important to evaluating exposure.

Sample Cumulative Distribution Function - a function that estimates the theoretical cumulative distribution function of a population. If a sample of n independent values is available, the value of the sample cumulative distribution at x is the proportion of the sample values that are less than or equal to x.

Standard Deviation - a measure of the variability of the values in a sample or a population. The positive square root of the variance of the distribution.

Statistical Inference - the process of using knowledge about samples to make statements about the population.

Statistical Significance - an inference that the probability of an observed pattern (with respect to the data being measured or the comparison being made) is so low that it is highly unlikely to have occurred by chance alone (within the constraints of the hypothesis being tested). The inference is that the hypothesis being tested is probably not true; that hypothesis is rejected in favor of a stated alternative hypothesis.

Statistically Selected Sample - a sample chosen based on a statistically valid sampling plan.

Stratified Random Sample - a sample obtained by separating the population elements into nonoverlapping groups called strata, and then selecting a simple random sample for each stratum.

Theoretical Cumulative Distribution Function - a function that uniquely defines the probability distribution of a random variable, x. The function specifies the probability that the random variable assumes a value less than or equal to x.

Worst Case - a semiquantitative term referring to the maximum possible exposure that can conceivably occur, whether or not this exposure actually occurs or is observed in a specific population.

APPENDIX A

SPREADSHEET MATRIX FOR TYPE 1
EXAMPLE DATA SET
FULL SHIFT
PERSONAL SAMPLES

APPENDIX A

The data set presented in Appendix A represent 516 full-shift personal samples grouped into 58 initial categories. In addition to these data, 37 short-term samples and 232 area samples were collected. Since these data were not used in the example analysis they were set aside and are not presented in this appendix.

APPENDIX B

BACKGROUND INFORMATION ON STATISTICAL METHODOLOGY

APPENDIX B

BACKGROUND INFORMATION ON STATISTICAL METHODOLOGY

This appendix presents background information for statistical methods used in these guidelines, as well as others that may be useful in the context of occupational exposure monitoring. Some of the topics include log-normal distributions, analysis of variance, data transformations, tests of distributions, cluster analysis, outliers, and confidence intervals. The engineer may wish to become familiar with these methods and the statistical assumptions associated with each method. References such as Massey, 51, for the K-S test; Cochran, 63; Daniel, 78; Conover, 80; etc. should be obtained and consulted, as needed. EPA statisticians should also be consulted, as required.

Box-and-Whisker Plot

Box-and-whisker plots are useful for the graphical identification of possible outliers. The box plot presents a clear depiction of outliers, compared to the majority of the whole data set. The box portion of a box plot extends from the 25th percentile to the 75th percentile of the observed data (i.e., 25% of the observations are at or below the bottom of the box and 25% are above the top of the box). That range is called the interquartile range. The whiskers extending from the box cover only 1.5 times the interquartile range. Any points outside 1.5 times the range are presented individually. This allows clear identification of outliers.

Analysis of Variance

Analysis of variance is the basis for many statistical techniques. It is applicable to normally distributed data (observations for which the errors are assumed to be normally distributed), especially in the context of testing for significance of possible explanatory variables.

Nested analysis of variance is a particular form of analysis of variance that addresses the issues associated with hierarchical (nested) data structures. In such structures, the variations induced by one variable are nested within (vary around) means that are dependent on the value of another variable, and which may also vary. Box (78) presents a nice discussion of nested designs and their analysis. Samuels (85) discusses the nested structure for occupational exposure data.

Tests of Distributions

The guidelines assume a log-normal distribution but there are three common approaches to quantitatively testing groups of data to determine if they can be described by certain distributions: the Shapiro-Wilk statistic, the Kolmogorov-Smirnov approach, and the ratio statistic.

The Shapiro Wilk statistic involves covariances between the order statistics of a standard normal distribution. It is similar to a test that examines the correlation (squared) between the observed order statistics and hypothetical order statistics. Order statistics are simply the observations (or hypothetical values) arranged in ascending order: the first order statistic is the smallest value, the second order statistic is the next smallest, etc. Simulation studies have suggested that the Shapiro-Wilk statistic is more powerful than the Kolmogorov-Smirnov test. Note that it can be applied only for testing for normality. Bickel (77) gives a short discussion and references to material on the Shapiro-Wilk statistic.

The ratio test was proposed in Waters (91) as a procedure for testing for log-normality. It makes use of two estimates of the mean of a log-normal distribution. In fact the ratio that gives this test its name is the ratio of those two estimates and is very easy to calculate. Its application requires the estimation of the coefficient of variation (related to the geometric standard deviation) and use of tables derived in Waters (91). Those tables are not complete for large values of the coefficient of variation. Waters (91) compared the ratio test favorably to the Shapiro-Wilk and Kolmogorov-Smirnov approach.

The Kolmogorov-Smirnov (K-S) approach is a widely used technique. The particular application presented here is for testing for normality, and has been called the Lilliefors test. K-S approaches are applicable more generally for testing for a variety of distributions.

The calculations needed to apply the Lilliefors test are discussed in some detail here. The procedure consists of the following: 1) deriving the sample cumulative distribution function for the observed data; 2) calculating the sample mean of the data (which may be concentrations if testing for normality or log-transformed concentrations if testing for log-normality); 3) calculating the sample standard deviation of the data; 4) standardizing the data; 5) determining the theoretical cumulative distribution; 6) identifying the value for passing the K-S test (the critical value); 7) calculating the maximum difference between the theoretical cumulative distribution and the sample cumulative distribution (the test statistic); and 8) determining if the data pass the test.

1. Derive the Sample Cumulative Distribution Function

The monitoring results for a group are arranged in ascending order: lowest value first and the highest value last. Next, the values for the sample cumulative distribution function are calculated on the sorted data. The cumulative distribution function for each data point is equal to the proportion of values less than or equal to the given point, as presented in Equation B1.

$$SCD_i = i / n \qquad \text{Equation B1}$$

where:

SCD_i = the sample cumulative distribution function value for observation i
n = number of data points.

2. Calculate the Sample Mean of the Data

The sample mean of the data is calculated using Equation 9 (for the concentrations) or Equation 2 (for transformed data) from Step 19.

3. Calculate the Sample Standard Deviation of the Data

The sample standard deviation of the transformed data is calculated using Equation 10 (for the concentrations) or Equation 3 (for transformed data) from Step 19.

4. Standardize the Data

The purpose of this step is to standardize the data to the standard normal distribution curve. The equation for standardizing the transformed data is presented in Equation B2.

$$z_i = (y_i - SM)/SSD \qquad \text{Equation B2}$$

where:

z_i = a standardized data point

SSD = the sample standard deviation of the data from 3 above
SM = the sample mean of the data from 2 above
y_i = a data point (either a concentration or transformed concentration)

Subtracting SM shifts the mean to zero, and then dividing by SSD scales the variable so that the standard deviation is 1 rather than SSD.

5. Determine the Theoretical Cumulative Distribution

This step consists of calculating a values corresponding to a theoretical (normal) cumulative distribution function for the standardized transformed data. The distribution may be calculated manually using a standard normal table or determined by one of several statistical software packages (see Appendix C). A standard normal table may be found in many statistical texts, including Bickel (77).

6. Identify the Value for Passing the K-S Test

Table B1 presents critical values for the Lilliefors test (Conover, 80).

The critical values depend on the sample size and the level of statistical significance required. For sample sizes between the values on Table B1, the value for the next highest sample size can be used.

7. Calculate the Differences Between the Values of the Theoretical Cumulative Distribution and the Sample Cumulative Distribution

This step consists of subtracting the values of the theoretical cumulative distribution function from the values of the sample cumulative distribution function and taking the absolute value, for each of the data points. The goal is to identify the maximum vertical difference between the sample and theoretical cumulative distribution functions. Since the sample cumulative distribution function is constant for values between the data points, the differences examined should include those between the value of the sample cumulative distribution function at a particular data point value and (1) the value of the theoretical cumulative distribution function at that data point value and (2) the value of the theoretical cumulative distribution function at the next data point value.

8. Determine If the Data Pass the Lilliefors Test

If none of the absolute values of the differences between the theoretical cumulative distribution and the sample cumulative distribution exceed the critical value identified in 6 above, then it may be concluded that the data can be described by a normal distribution. If one or more of the absolute differences exceed the critical value, the normal distribution is not appropriate.

TABLE B1. CRITICAL VALUES FOR LILLIEFORS TEST (Conover, 80)

	Level of significance				
Sample size	0.20	0.15	0.10	0.05	0.01
4	0.300	0.319	0.352	0.381	0.417
5	0.285	0.299	0.315	0.337	0.405
6	0.265	0.277	0.294	0.319	0.364
7	0.247	0.258	0.276	0.300	0.348
8	0.233	0.244	0.261	0.285	0.331
9	0.223	0.233	0.249	0.271	0.311
10	0.215	0.224	0.239	0.258	0.294

11	0.206	0.217	0.230	0.249	0.284
12	0.199	0.212	0.223	0.242	0.275
13	0.190	0.202	0.214	0.234	0.268
14	0.183	0.194	0.207	0.227	0.261
15	0.177	0.187	0.201	0.220	0.257
16	0.173	0.182	0.195	0.213	0.250
17	0.169	0.177	0.189	0.206	0.245
18	0.166	0.173	0.184	0.200	0.239
19	0.163	0.169	0.179	0.195	0.235
20	0.160	0.166	0.174	0.190	0.231
25	0.142	0.147	0.158	0.173	0.200
30	0.131	0.136	0.144	0.161	0.187
Over 30	$\frac{0.736}{\sqrt{N}}$	$\frac{0.768}{\sqrt{N}}$	$\frac{0.805}{\sqrt{N}}$	$\frac{0.886}{\sqrt{N}}$	$\frac{1.031}{\sqrt{N}}$

Data Transformations

The guidelines consider a simple data transformation, the log transformation. That transformation is just one of a family called the Box and Cox transformations. Such transformations are often considered prior to analysis of data in order to make the data more normal and to make the variances in different groups more similar, both of which are desirable for most analysis of variance approaches, for example. The reader is referred to Stoline (91) for a discussion of the Box and Cox family of transformations applied to environmental data. Samuels (85) also considers transformations other than the log transformation for occupational exposure data. The guidelines do not recommend transformations other than the log transformation because of the computations involved, because the properties (e.g., mean and standard deviation) of the log-normal distribution are well known whereas the interpretation and calculation of descriptive statistics based on other transformations is not straightforward, and because the log-normal distribution is an accepted distribution for concentration data.

Log-normal Distribution

The log-normal distribution has been studied and applied to concentration data for many years (Aitchison, 57; Johnson, 70). The estimation of the mean of the log-normal distribution is discussed in detail in Attfield (92). Note that the formula for MLE_A in Attfield (92) is incorrect as stated: multiply the formula given by exp(x) to get the corrected value for MLE_A. Confidence limits for the mean of a log-normal distribution are presented in Armstrong (92). Samuels (85) shows how confidence intervals for the concentration data means can be derived from standard deviations and standard errors associated with transformed data.

Confidence Intervals

The calculation of confidence interval is an important means of presenting the degree of certainty about the estimates of any particular parameter. It is important to note that a confidence interval for a mean, for example, must be based on the variance associated with that estimate, not with the variance associated with the individual observations in the population. Thus, the standard error of the mean (which is the square root of the variance of the mean estimator) should be used to define a confidence interval for the mean.

Confidence intervals for means also depend on the data structure and the distribution of the data. Although asymptotically (as the sample size gets very large) a mean will be normally distributed, no matter what the underlying distribution of the observations may be, for relatively small sample sizes the normal approximation may be poor. Thus, confidence intervals for a log-normal mean, for example, have been specifically defined (Armstrong, 92). Standard errors and therefore confidence intervals can be defined for transformed concentrations and converted back to the original scale (Samuels, 85). Standard errors that take into account nested data structures can also be computed (Samuels, 85) and used to define confidence intervals.

Techniques to Combine Groups

One of the final quantitative steps in the guidelines is to obtain statistics for combinations of groups. As is discussed in the text, this should only be attempted when appropriate. The only techniques identified as appropriate are from stratified sampling theory. These techniques can be considered because they allow for estimation of means and standard deviations across groups with widely different population sizes. The properties of these estimates are not known for nonrandom sampled data. This fact should be stated if such estimates are used.

Cluster Analysis

Another approach to defining groups for statistical analysis is based on a procedure known as cluster analysis. That approach examines characteristics of the measurements within groups (clusters) and determines when two groups are similar enough to be combined. The cluster analysis approach is described here is some detail.

Cluster analysis is an iterative procedure by which clusters are combined. Combination proceeds in order of similarity: the most similar groups are combined first, then the next most similar, etc. Each group of measurements (e.g., a set of observations sharing the same values for all the important exposure parameters identified by the engineer or industrial hygienist) starts out as a single cluster; when two groups are combined, the combined group replaces the two groups that were combined, for the purposes of comparison with other groups and additional combination.

In order to conduct a cluster analysis, some measure of similarity is required. The simplest measure, and one that can easily be used for routine application to occupational exposure data, is based on the mean values of the measurements within groups: two groups are considered similar when the difference between their mean values is small. This clustering method is referred to as the unweighted pair-group method using arithmetic averages (UPGMA).

The advantage of this method of clustering is that it does not require the specification or assumption of an underlying distribution for the measurements within the groups. A disadvantage is that this method only compares the mean values within groups and does not consider other descriptors of the within-group measurements, such as variation. Some other methods for defining the similarity of groups are discussed and compared with the UPGMA method in the SAS manual. In some applications those other methods may be more appropriate than the simple UPGMA procedure. Consultation with a statistician is recommended in those cases, and may even be required when the UPGMA method is all that is desired.

A cluster analysis can proceed until all the groups are combined into one cluster. Output from a computer package will specify which clusters are combined at each step and the similarity (difference in means for the UPGMA method) of the clusters combined at each step. The engineer can examine the output and determine at what point the clustering is sufficient, where "sufficient" clustering is based on consideration of sample sizes attained, on the similarity of the clusters that are combined, or on a combination of those two factors.

The goal of this procedure is to increase sample sizes and define uniform groups. It is inappropriate to combine groups that are quite dissimilar, just to get big sample sizes. Thus, some decision by the engineer, in consultation with the statistician, must be made about the weight to be given to the conflicting pressures of those two considerations (sample size vs uniformity). It is recommended that the engineer and statistician decide on a "stopping rule" prior to the running of the cluster analysis. The stopping rule will specify the largest measure of similarity (largest difference in means for the UPGMA method) that will be considered acceptable for combination to occur. The knowledge of the engineer and the statistician is required to select a stopping rule, as there is currently no statistical test or probabilistic measure that can tell the user when the clustering of groups is inappropriate. An examination of the initial groups, their means, and the overall mean for all groups may provide some indication of a stopping rule to consider.

One drawback to the cluster technique is that it can combine groups which do not belong together from an engineering perspective. *A priori* selection of appropriate and inappropriate groupings of data based on engineering judgement can be used to prevent inappropriate clustering of the data. The ANOVA technique discussed in Step 15 does not have this problem. However, the ANOVA technique is most appropriate for data from designed, controlled experiments.

APPENDIX C

LISTING OF COMPUTER SOFTWARE FOR VARIOUS STATISTICAL ANALYSES

APPENDIX C

LISTING OF COMPUTER SOFTWARE FOR VARIOUS STATISTICAL ANALYSES

Box-and-Whisker Plot

There are many software packages available on the PC for this technique. These include CSS, NWA Statpak, Solo, SPSS/PC Plus, Statgraphics, Statpac Gold, Systat/Sygraph, SAS, and BMDP.

Analysis of Variance

Analysis of variance is a standard statistical tool available in the software packages CSS, NWA Statpak, Solo, SPSS/PC Plus, Statgraphics, Statpac Gold, Systat/Sygraph, SAS, and BMDP. Not all of these packages can provide the results needed to obtain variance components for a nested analysis of variance. SAS has a special procedure, PROC NESTED, which does just that.

Distribution Tests

The Shapiro-Wilk test is provided as an option in the SAS procedure PROC UNIVARIATE.

Many software statistical packages have the K-S type test procedures for the PC: CSS, NWA Statpak, SPSS/PC Plus, Statgraphics, Statpac Gold, and Systat/Sysgraph. For these packages, the user compares a normal distribution to a set of data.

Theoretical Cumulative Distribution

The many software packages available for computing the standard normal theoretical cumulative distribution function include CSS, NWA Statpak, Solo, SPSS/PC Plus, Statgraphics, Statpac Gold, Systat/Sysgraph, SAS, and BMDP.